# Digital Cognitive Assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report

*Eliza Passell[1], Daniel G. Dillon[2.3], Justin T. Baker[1,2], Sarah C. Vogel[1], Luke S. Scheuer[1], Nicholas L. Mirin[1], Lauren A. Rutter[1,2], Diego A. Pizzagalli[2,3], Laura Germine*[1,2]*

1. *Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA*
2. *Department of Psychiatry, Harvard Medical School, Boston, MA, USA*
3. *Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA, USA*

**Corresponding author:*
*Laura Germine, PhD*
*Address: McLean Hospital, 115 Mill St, Belmont, MA 02478, USA*
*Email: lgermine@mclean.harvard.edu*

**Acknowledgments:** This report was commissioned by the NIH National Institute of Mental Health as part of contract HHSN271201700776P (PI: Laura Germine). We would like to thank Jasmine Kaduthodil, Claire Peterson, Emily Wu, Alison Tassone, Olivia Nicholson, and Dana Zeng for their help putting this report together.

## Abstract

*Background:* Digital technology has become a major target area for the development of assessments that can be deployed through mobile devices, across large cohorts, and in naturalistic environments. Here, we summarize the results of a report commissioned by the National Institute of Mental Health (HHSN271201700776P) to evaluate mobile assessments of cognition and their appropriateness for deployment in a field test battery.

*Method:* Using data from over 100,000 participants tested through our digital research platform, TestMyBrain.org, we analyze the appropriateness of 25 standard tests of cognition and information processing for field test battery use. Measures are evaluated in terms of their psychometric properties, validity, engagement, and sensitivity to variations in device hardware and software. We also define a minimum duration for acceptable reliability (minDAR) across all 25 tests, operationalized as the duration a test needs to be to achieve an internal reliability of at least 0.7 for primary outcome measures.

*Results:* We note that many tests adapted from experimental approaches, particularly those involving aspects of positive and negative valence, need further development to achieve acceptable length and reliability (based on very high minDAR values, e.g. 180 minutes vs 3 minutes for threat biases in memory vs. memory alone). Device variability also presents a confound for reaction time tests (e.g. iOS vs Android Cohen's $d = 0.4$ for simple reaction time, $p < 0.001$). Areas of focus for development of such measures are described.

*Conclusion:* Digital cognitive assessment is a promising methodology for large sample studies at relatively low cost. There are notable areas where further research and development work is needed, however, to fully realize the potential for cognitive phenotyping at scale.

# TABLE OF CONTENTS

## Overview

For this report, and at the request of the NIH National Institute of Mental Health Research Domain Criteria Unit (NIMH RDoC Unit), we evaluated 21 cognitive assessments that were developed for administration through *TestMyBrain.org* with respect to their psychometric, technology, and engagement characteristics. The results of this evaluation are summarized as individual test reports in the pages that follow as well as in a single table that follows this summary. The goal of this analysis was not to evaluate or comment on the theoretical motivation for inclusion of any particular test or particular construct in a field test battery, but rather to give the NIMH RDoC leadership the information necessary to evaluate whether any particular test might be suitable for translation to field test use in the near term. As such, the reports comment minimally on the scientific justification and background for any particular test, but rather focus on analysis and evaluation of data collected through *TestMyBrain.org* to determine how well each test meets a set of predefined criteria, specifically: good psychometric properties, suitability for measuring RDoC constructs, evidence for validity, ease of administration in a field test setting, device invariance, participant tolerability / engagement, as well as any recommendations for modification to make a test appropriate for field test use. The table that follows gives a general test-by-test summary of the results of the evaluation, with a few points highlighted below.

1. **Tests derived from experimental traditions such as cognitive psychology and cognitive neuroscience take substantially more resource investment to adapt to a field test setting.** Most tests derived from standard neuropsychological tests - where scalability and standardization of methods is a major goal - are readily adaptable for delivery on mobile devices as long as the general procedure did not inherently rely on things like drawing, writing, or free recall. Tests derived cognitive neuroscience and experimental psychology, on the other hand, varied substantially in how easily they were adaptable to a field test setting. As these tests were developed to maximize stimulus control, precision of response measurement, and tend to rely on many trials, field test versions of these tests tended to be more burdensome for participants, less reliable, and more difficult to adapt to a large range of devices. We observe, for example, that neither of our tests of negative valence (aside from judgement of angry or fearful faces) produced data with acceptable psychometric characteristics. On the other hand, it is certainly possible to build high quality tests modeled on paradigms from cognitive neuroscience and experimental psychology -- but these take significantly more resource investment than adaptations of standard neuropsychological tests before appropriate field test assessments can be delivered. Our sixth significant iteration of the probabilistic reward test, for example, produces reliable response biases in a relatively short period of time.This is an area that needs further resource investment if field test versions of popular mechanism-based tests derived from experimental psychology and cognitive neuroscience are to be used.

2. **Brief and reliable performance-based tests of positive and negative valence represent a major gap for an “in the field” approach to the RDoC.** Related to the point above -- and

a subject to discussion in other efforts to define consensus or recommended batteries for use in large cohorts and on mobile / internet-connected devices - there are very few good measures of some of the fundamental RDoC constructs that are appropriate for a field test battery. If the NIMH seeks to build a taxonomy based on RDoC domains that depends on behavioral measurement in large cohorts, the lack of such measures will pose a major barrier to progress.

3. **Device variance poses a significant threat to the validity of measures that depend on reaction time latency.** Tests where average response times were a major outcome showed significant variability across devices. This is unfortunate as these tests also produce the most reliable scores in the shortest period of time. The shorter the average reaction time measured, the more device variability was an issue - e.g. simple reaction time was substantially affected compared to other tests with longer average reaction times. We estimate, for example, that differences between Android phones and desktop / laptop computers for a test like simple reaction time produce differences comparable to the differences between performance in someone age 20 vs age 70 (Cohen's d = 0.65). Comparing simple reaction time scores on Android vs iOS phones can produce differences equivalent to differences between individuals with major depressive disorder (and associated psychomotor slowing) vs. healthy controls (Cohen's d = 0.4). These differences are *after* correction for the expected effects of demographic variables, as estimated by vocabulary score differences by device. Approaches to dealing with such variation in response time latency between devices needs to be addressed in any designs using these tests (Germine, Reinecke & Chaytor, 2019).

4. **Better psychometric properties = better participant engagement.** Average participant ratings for a particular test and the test's internal reliability were highly correlated (r = 0.54). Participant retention and test length were also closely highly correlated (r = -0.53). This indicates that short tests with good psychometric properties will have the highest engagement and satisfaction. We provide a metric (minDAR = minimum duration for acceptable reliability) that allows tests to be directly compared in terms of how long it takes to yield the same amount of information. Given these characteristics, in general, tests with lower minDAR values will be better for participant engagement and maximizing the chance that participants will complete a test.

A summary of our evaluation of individual tests is included in the pages that follow, starting with a table summarizing these results and followed by individual test reports, in alphabetical order.

### Summary of Measures

| Task Name | Psychometrics | Internal Reliability | RDoC-iness | Validity |
|---|---|---|---|---|
| TMB Matrix Reasoning | Excellent | 0.89 | Low | High |
| TMB Threat / Neutral Sternberg Memory Test | Poor | 0.14 | High | Med |
| TMB / TAU Threat / Neutral Dot Probe Test | Very Poor | 0 | High | Low |
| TMB / Baron-Cohen Reading the Mind in the Eyes test | Good | 0.78 | High | High |
| TMB Probabilistic Reward Test | Excellent | 0.85 | High | not enough evidence for current version |
| TMB Flicker Change Detection | Good | 0.78 | Med | High |
| TMB Gradual Onset Continuous Performance Test | Good | 0.78 | High | High |
| TMB Multiracial Face Emotion Identification Test | Good | 0.75 | High | High |
| TMB Delay Discounting | Excellent | 0.92 | High | Med |
| TMB Flanker Test | Good | 0.77 | High | Med |
| TMB Choice Reaction Time Test | Excellent | 0.95 | Med | Med |
| TMB Digit Symbol Matching Test | Excellent | 0.93 | Low | High |
| TMB / Dillon Emotional Word Memory Test | unknown | unknown | High | not enough evidence for current version |
| TMB Multiple Object Tracking | Excellent | 0.92 | Low | High |
| TMB Visual Paired Associates | Good | 0.79 | Med | Med |
| TMB Verbal Paired Associates | Excellent | 0.82 | Med | Med |
| TMB Simple Reaction Time | Excellent | 0.93 | Low | High |
| TMB Anger Sensitivity Test | Good | 0.78 | Med | Med |
| TMB Happiness Sensitivity Test | Good | 0.7 | Med | Med |
| TMB Fear Sensitivity Test | Good | 0.8 | Med | Med |
| TMB Synonym Vocabulary | Good | 0.83 | Low | High |

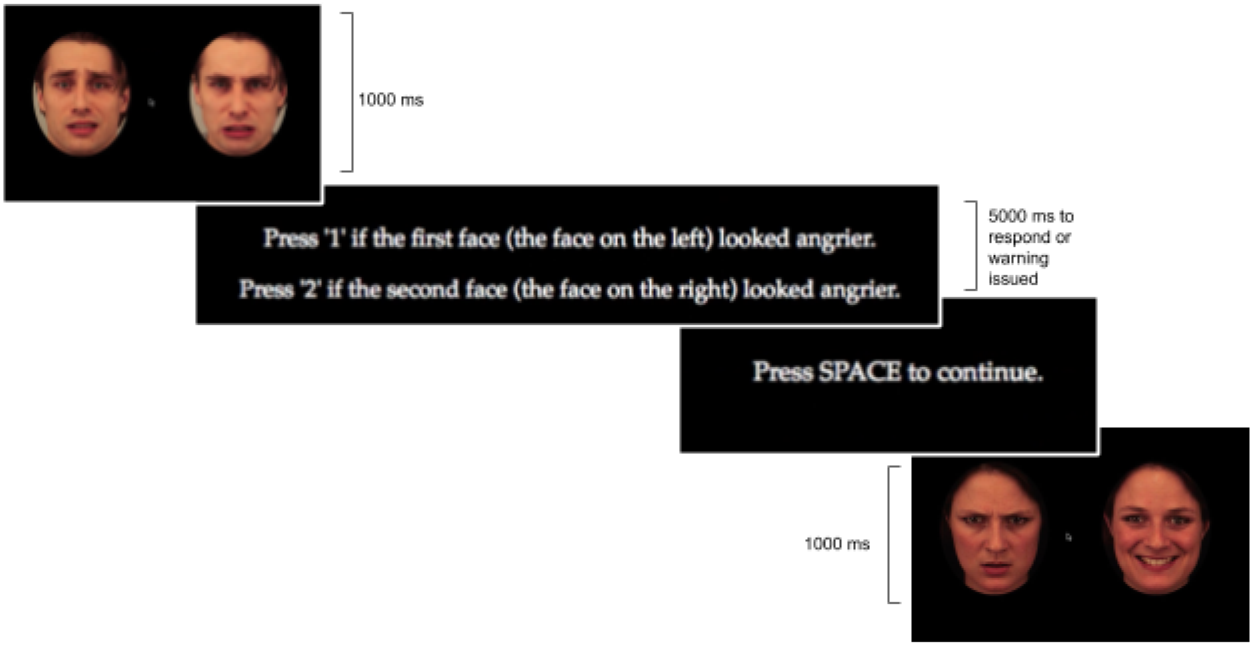
| Task Name | **Ease of Administration in Field Test Setting** | **Device Invariance** | **Length (minutes)** | **Minimum Duration for Acceptable Reliability (minDAR)** |
|---|---|---|---|---|
| TMB Matrix Reasoning | High | High | 8.5 | 3 |
| TMB Threat / Neutral Sternberg Memory Test | Med (confusing) | unknown (too unreliable) | 12 | 180 |
| TMB / TAU Threat / Neutral Dot Probe Test | High | unknown (too unreliable) | 4.3 | Inf |
| TMB / Baron-Cohen Reading the Mind in the Eyes test | Med (vocab demands) | High | 10 | 7 |
| TMB Probabilistic Reward Test | High | unknown (not enough data) | 9.5 | 4.5 |
| TMB Flicker Change Detection | Med (difficulty responding) | Med (difficulty responding) | 5 | 3.5 |
| TMB Gradual Onset Continuous Performance Test | High | High | 7 | 4.9 |
| TMB Multiracial Face Emotion Identification Test | High | High | 2.5 | 2 |
| TMB Delay Discounting | Med (confusing) | High | 5 | 1 |
| TMB Flanker Test | Med (perception challenges) | High | 5.5 | 3.9 |
| TMB Choice Reaction Time Test | Med (confusing) | Med (RT latency) | 2.5 | 0.5 |
| TMB Digit Symbol Matching Test | High | Med (RT latency) | 3 | 0.5 |
| TMB / Dillon Emotional Word Memory Test | High | unknown (not enough data) | 17 | unknown |
| TMB Multiple Object Tracking | High | High | 10 | 2 |
| TMB Visual Paired Associates | High | High | 5 | 3.5 |
| TMB Verbal Paired Associates | High | High | 5 | 3 |
| TMB Simple Reaction Time | High | Low (RT latency) | 1.5 | 0.5 |
| TMB Anger Sensitivity Test | High | High | 3.5 | 2.5 |
| TMB Happiness Sensitivity Test | High | High | 3.5 | 3.5 |
| TMB Fear Sensitivity Test | High | High | 3.5 | 2.5 |
| TMB Synonym Vocabulary | High | High | 4 | 2 |

*retention for memory tests not comparable to other measures due to inclusion of interstitial delay tasks

| Task Name | Participant Retention / Burden | Participant Ratings / Satisfaction | Appropriate for Field Test Use | Recommended Changes |
|---|---|---|---|---|
| TMB Matrix Reasoning | 90% | 3.8 | Yes | IRT Adaptive |
| TMB Threat / Neutral Sternberg Memory Test | 61% | 3.7 | No | -- |
| TMB / TAU Threat / Neutral Dot Probe Test | 75% | 3.1 | No | -- |
| TMB / Baron-Cohen Reading the Mind in the Eyes test | 76% | 3.7 | Yes | Multiracial Version |
| TMB Probabilistic Reward Test | 38% | 3.5 | w/modifications | Reduce length |
| TMB Flicker Change Detection | 86% | 3.7 | Yes | Reduce stimulus density |
| TMB Gradual Onset Continuous Performance Test | 64% | 3.6 | Yes | Reduce length |
| TMB Multiracial Face Emotion Identification Test | 97% | 3.7 | Yes | None |
| TMB Delay Discounting | 87% | 3.5 | Yes | Reduce length |
| TMB Flanker Test | 69% | 3.6 | Yes | Visual perception screen |
| TMB Choice Reaction Time Test | 88% | 3.9 | Yes | None |
| TMB Digit Symbol Matching Test | 93% | 4 | Yes | Alternate forms needed |
| TMB / Dillon Emotional Word Memory Test | 43%* | 3.5 | Yes | Reduce length |
| TMB Multiple Object Tracking | 85% | 3.9 | Yes | Reduce length |
| TMB Visual Paired Associates | 56%* | 4 | Yes | Alternate forms needed |
| TMB Verbal Paired Associates | 43%* | 3.8 | Yes | None |
| TMB Simple Reaction Time | 91% | 3.85 | Yes | Reduce length |
| TMB Anger Sensitivity Test | 67% | 3.4 | Yes | Multiracial Version |
| TMB Happiness Sensitivity Test | 87% | 3.4 | Yes | Multiracial Version |
| TMB Fear Sensitivity Test | 67% | 3.4 | Yes | Multiracial Version |
| TMB Synonym Vocabulary | 83% | 3.85 | Yes | IRT Adaptive |

### TMB Anger Sensitivity

Constructs Measured: potential threat, social communication/reception of facial communication, understanding mental states
Duration: 3.3 minutes
Sample size for which normative data are available: 16,913
Demo Link: [http://www.testmybrain.org/tests/emotion_comparison/anger.html](http://www.testmybrain.org/tests/emotion_comparison/anger.html)
Description of procedure: Judge which of two faces is more angry. Faces will flash on the screen for a short period of time and participants are instructed to determine which of the two faces displays an angrier expression.



Press '1' if the first face (the face on the left) looked angrier.

Press '2' if the second face (the face on the right) looked angrier.

Press SPACE to continue.

This task assesses sensitivity to differences in anger intensity, independent of response bias and differences in emotion identification or categorization (Rutter et al., 2019). Advantages of the task is it allows issues related to categorization, verbalization, response bias to be dissociated from sensitivity to specific face emotions. It is also short and easy to administer across a range of mobile device types. Disadvantages are that the task is not yet validated with respect to clinical conditions or psychopathology and is considered burdensome by participants despite its relatively short length.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Emotion sensitivity or emotion comparison tests are not included in the RDoC Council Workgroup Report on Behavioral Assessments, so we consider this test **PRIORITY 2.**

### Current Applications

All three TMB Emotion Sensitivity tests are being used and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

The primary outcome measure from this test is accuracy, based on proportion or number correct out of 56 trials. This score reflects the participant's ability to detect differences in happiness between faces. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures is less clear.

This test shows good reliability, especially given test length. Internal reliability (split-half) of accuracy is 0.78, as calculated from a sample of 5000 participants who completed this test on TestMyBrain.

Sociodemographic effects were estimated based on the accuracy scores of the 14,008 participants for whom demographic data was available. This population had a mean age of 26.53 and was 58.17% female. The distribution of scores is relatively normal with some ceiling effects (see Figure 1A). Performance is variable across the lifespan; scores increase during adolescence and plateau throughout adulthood before declining after age 50 (see Figure 1B). Female participants show slightly higher performance than male participants (see Figure 1C). Performance increases slightly with level of education, though this effect is not consistent between more highly educated participant groups (see Figure 1D).

This test does not show evidence of practice effects. First-time participants have a mean score of 46.37, while repeat participants have a mean score of 45.05.

### Validation

Performance on this test is correlated with other tests of emotion perception. It shows high correlation with performance on analogous tests of perception of happiness (r = 0.46, N = 12568, 95% CI [0.45, 0.47]) and fear (r = 0.53, N = 12312, 95% CI [0.52, 0.54]). All correlations have been adjusted for age. Scores are not associated with current anxiety as measured by the GAD-7 (r = -0.021, N = 531, 95% CI [-0.11, 0.064]), but are associated with depression symptoms as measured by the Beck Depression Inventory (r = -0.10, N = 486, 95% CI [-0.19, -0.015]).

### Appropriateness for Field Test Use

Before beginning the test, each participant completes 2 easy practice trials, which include immediate feedback and are repeated if the participant answers incorrectly. Therefore, difficulties in understanding the task should not present a barrier to completion.

*Device Effects.* Participants who took this test using mobile devices showed slightly lower performance than those who used laptop or desktop computers (iPhone mean = 46.91, SD = 5.07, N = 1192; iPad mean = 46.53, SD = 5.26, N = 517; Macintosh laptop/desktop mean = 48.70, SD = 4.54, N = 1133). Device type may have an impact on performance on this test; for instance, although comparable scores between iPhone and iPad suggest that screen size does not explain this difference, but may instead be due to differences in demographics or environmental context.

*Participant Burden.* This test is considered burdensome by participants. The mean participant rating for batteries on TestMyBrain containing this test is 3.40, compared to a site-wide mean participant rating of 3.7. Completion rates are 67.4%, which compares unfavorably with other tests on TestMyBrain.org (81% completion average).

## Further Development

The current version of this test relies on faces taken from predominantly Caucasian face databases, so the major limitation of this test is its use in diverse cohorts.  Versions of the test that include multiracial faces are recommended for broader applications.

Figure 1A. Distribution of Scores

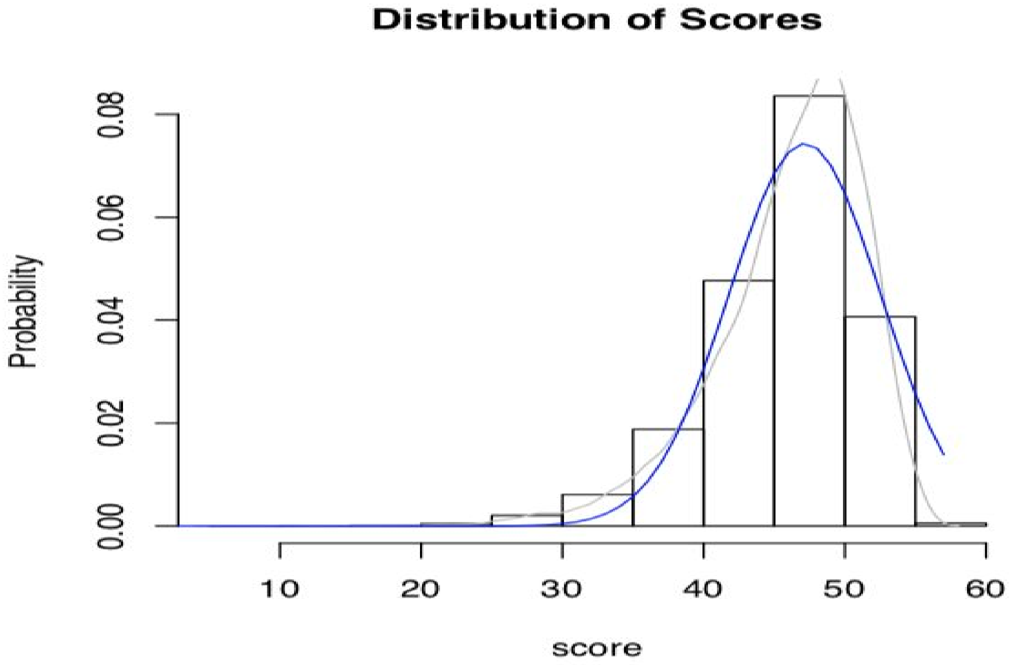


Figure 1B. Age-Related Differences in Performance

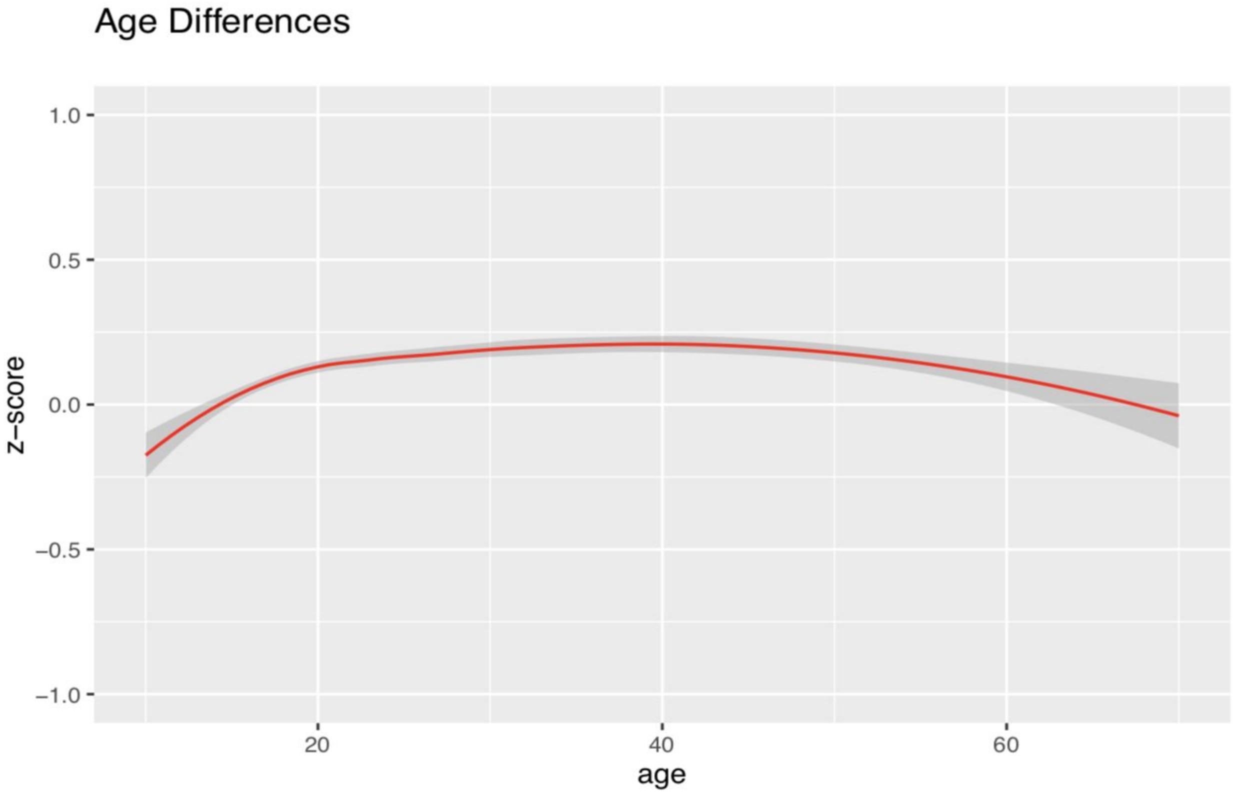


Figure 1C. Sex Differences in Performance

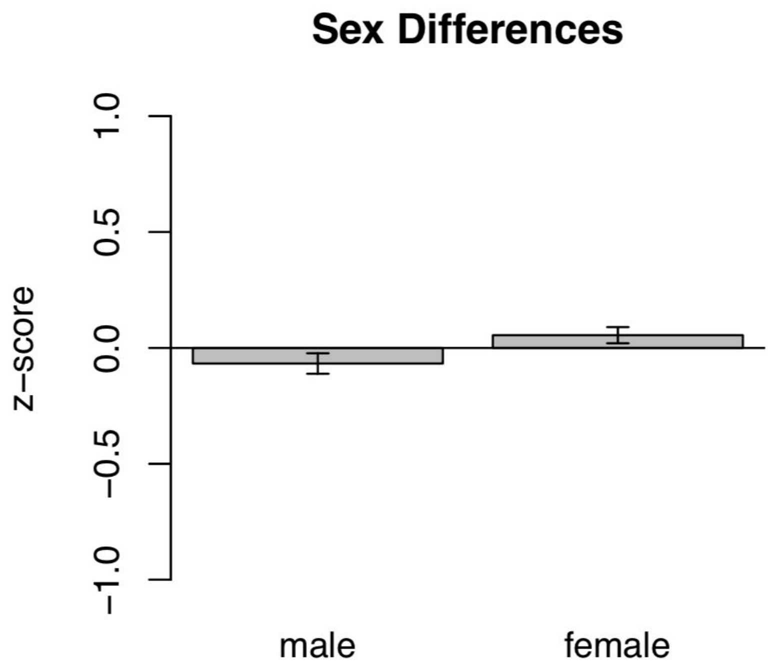
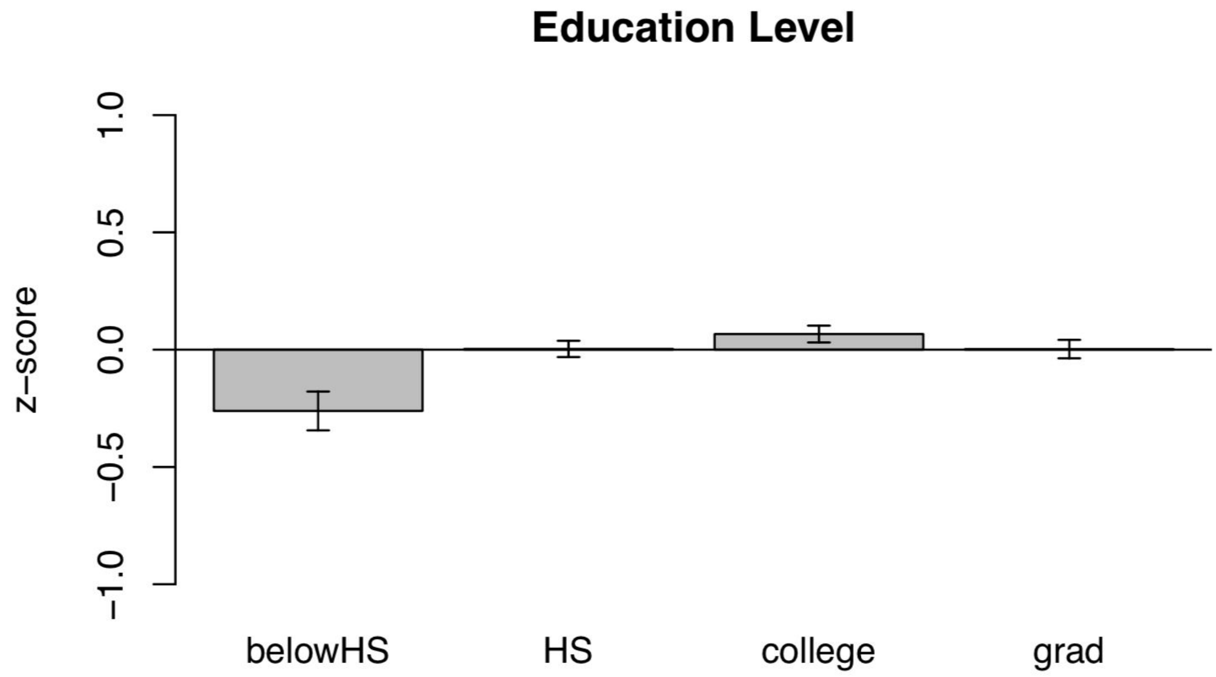
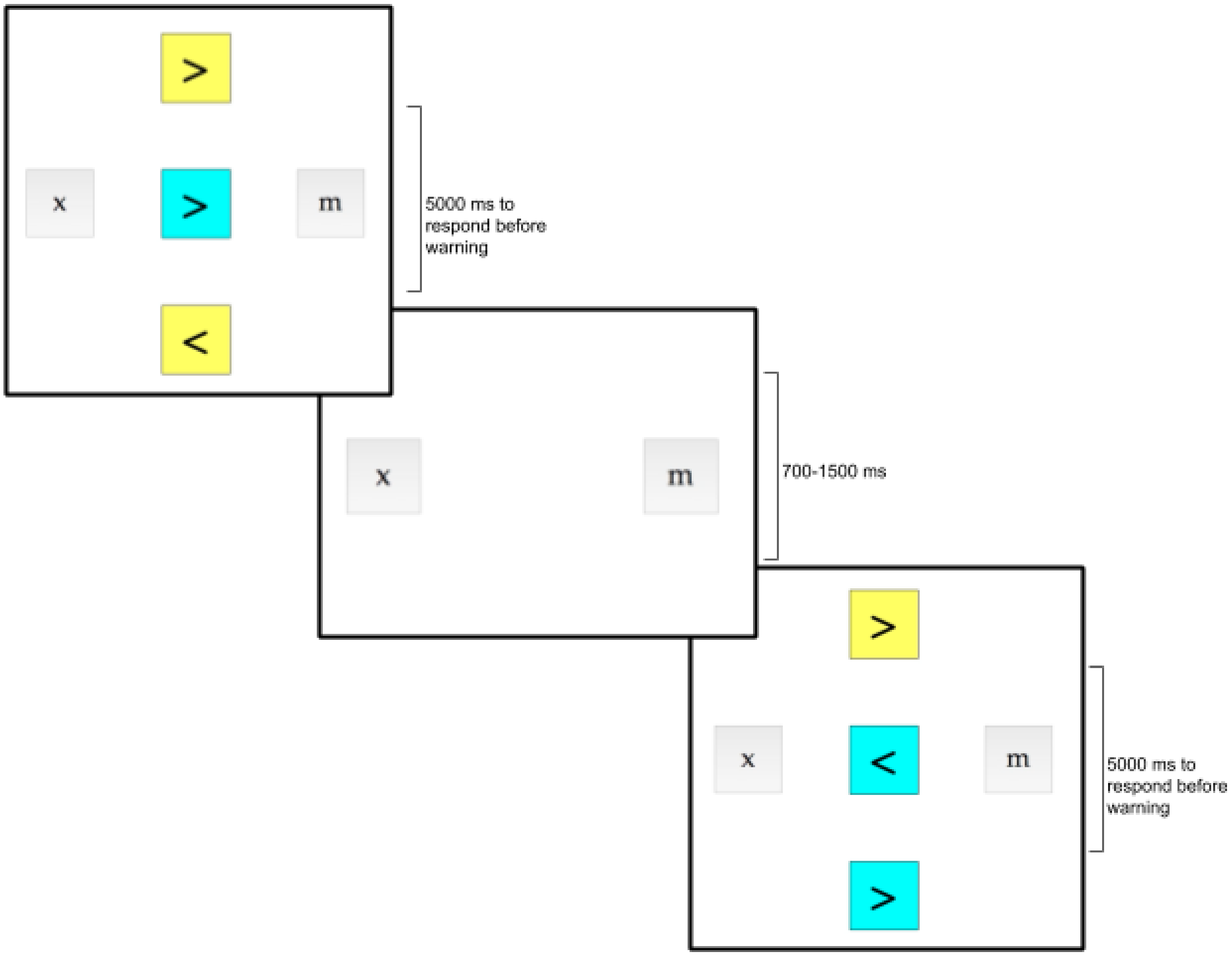


Figure 1D. Education-Related Differences in Performance



**TMB Choice RT**

Constructs Measured: processing speed, cognitive inhibition, cognitive control
Duration: 2.5 minutes
Sample size for which normative data are available: 18,556
Demo Link: http://www.testmybrain.org/tests/ChoiceRT/ChoiceRT.html
Description of procedure: Indicate the direction of an arrow that is a different color from the rest.


>
x
>
m
<
5000 ms to respond before warning
x
m
700-1500 ms
>
x
<
m
>
5000 ms to respond before warning

This is a standard format choice reaction time task, which requires a participant to efficiently select from multiple competing response options (Maljkovic & Nakayama, 1994). Advantages are that the task is very short, can be completed across a range of mobile devices, and is enjoyable to participants. Disadvantages are that this particular format of a choice reaction time test has not been validated with respect to clinical conditions or psychopathology. The procedure can sometimes be confusing to participants as well, who do not pay adequate attention during practice trials.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Although this is not technically a Flanker test, the response conflict and cognitive control demands are very similar to a flanker test. Flanker tests are included in the RDoC Council Workgroup Report on Behavioral Assessments, however, so this task is designated **PRIORITY 1.**

### Current Applications

The TMB Choice Reaction Time test is currently included in several major initiatives, including as part of the NIMH Aurora study and as part of the Broad Neuropsychiatric Phenotyping Initiative. Translation of the test into standard Chinese and Spanish is currently being funded by the Broad Institute.

### Psychometric Characteristics

The Choice RT test measures both reaction time and accuracy. The main outcome calculated from this test is median reaction time for correct trials, or median reaction time corrected for accuracy (inverse efficiency score: median RT / proportion correct) where speed accuracy trade-offs are a concern. For participant feedback purposes, reaction times are transformed (10000 / rt) to yield a number that typically ranges between 1 and 25 that corresponds to “speed”. All analysis in this report has been completed using both median reaction time on correct trials and inverse efficiency score.

This test has shown excellent reliability; for median reaction time on correct trials, internal reliability (split-half) was 0.95 (calculated using the subset of the participant pool enrolled through the Aurora project, n = 617). Internal reliability (split-half) for inverse efficiency score was 0.87 (calculated from 5000 participants who completed the test on TestMyBrain).

Sociodemographic effects were estimated based on median reaction times on correct trials for the 15,409 participants for whom demographic data was available. This population has a mean age of 29.9 and is 46.4% female. The distribution of scores is normal (see Figure 2A). Reaction time is variable across the lifespan, decreasing throughout adolescence, peaking in speed at approximately age 20, and increasing throughout adulthood (see Figure 2B). This pattern is typical of tests that measure processing speed. Male participants showed slightly faster reaction times than female participants (see Figure 2C). Reaction time decreases with education (see Figure 2D).

Inverse efficiency score showed similar sociodemographic effects as median reaction time. Scores were relatively normally distributed, with a small number of participants showing unusually high scores (see Figure 2E). Inverse efficiency score is variable across the lifespan and shows a similar pattern to median reaction time (see Figure 2F). Male participants showed slightly lower IES (indicating better performance) than female participants (see Figure 2G). IES decreases with increased education (see Figure 2H).

This test may have small practice effects; first-time participants had a mean median reaction time of 921.35, while repeat participants had a mean median reaction time of 867.17. These practice effects are also apparent in inverse efficiency scores: first-time participants have a mean IES of 1005.12, while repeat participants have a mean IES of 943.2118.

### Validation

Median reaction time on the Choice RT test correlate with performance on other tests measuring cognitive processing speed, cognitive control, and cognitive inhibition. It is correlated with simple reaction time (r = 0.40, n = 11178, 95% CI [0.38, 0.41]) and digit symbol matching, another task requiring quick responses to visual processing tasks (r = -0.41, n = 12397, 95% CI [-0.42, -0.39)). It is more modestly correlated with other tests of general cognitive ability, such as vocabulary (rho = -0.15, n = 549, 95% CI [-0.23, -0.070]), but not with tests of entirely distinct domains such as emotion recognition (r = -0.078, N = 529, 95% CI [-0.17, 0.0014]). This suggests that this test is able to specifically measure cognitive processing speed as a distinct faculty, separate from other cognitive abilities (Lee & Chabris, 2013).

Inverse efficiency score shows similar correlation with other tests, indicating that both metrics are of comparable validity. It is correlated with digit symbol matching (Spearman's rho = -0.46, n = 12397, 95% CI [-0.48, -0.45]) and simple reaction time (Spearman's rho = -0.43, n = 11178, 95% CI [-0.45, -0.42]). It is also correlated to a lesser extent with vocabulary (Spearman's rho = -0.34, n = 383, 95% CI [-0.42, -0.24]). Unlike the median reaction time on this test, inverse efficiency score is slightly correlated with performance in emotion recognition (r = -0.11, n = 370, 95% CI [-0.21, -0.0085]).

### Appropriateness for Field Test Use

This test is brief and well-tolerated by participants. To ensure that participants understand the task, the test includes a series of 4 practice trials before test trials begin. This ensures that scores and completion rates are not affected by participants' difficulty in understanding the requirements of the test. With these practice trials in place, there are minimal barriers to completion.

*Device Effects:* The Choice RT test is easy to administer across a wide variety of device types. However, since this test measures cognitive processing speed using reaction time, differences in device performance (such as device latency in registering input) are likely to impact measured scores. Our data showed that participants using desktop or laptop computers had lower median reaction times than those using mobile devices (iPhone mean: 931.70, SD = 2.53.47, N = 1635; iPad mean = 981.04, SD = 279.60, N = 987; Macintosh desktop/laptop mean = 876.16, SD = 271.51, N = 2303**).** Differences in device latency likely have a modest impact on median reaction time. Inverse efficiency score shows a similar pattern of device effects (iPhone mean = 998.15, SD = 345.51; iPad mean = 1048.85, SD = 371.03; Macintosh desktop/laptop mean = 948.96, SD = 358.42).

*Participant Burden:* The Choice RT test poses a low burden to participants. The mean participant rating for batteries containing this test was 3.9 out of 5, compared to an average of 3.7 for all batteries hosted on Test My Brain. 88.1% of participants who began this test completed it, which is substantially higher than sitewide completion of 81%.

### Further Development

This test can be readily modified for ecological momentary assessment designs and returns reliable scores from brief testing. It is possible that differences between devices used to take this test may affect the measurement of scores, so when using this test it would be necessary to control for the devices used by participants. Otherwise, this test is ready for field test use.

Figure 2A. Distribution of Scores (Median Reaction Time)

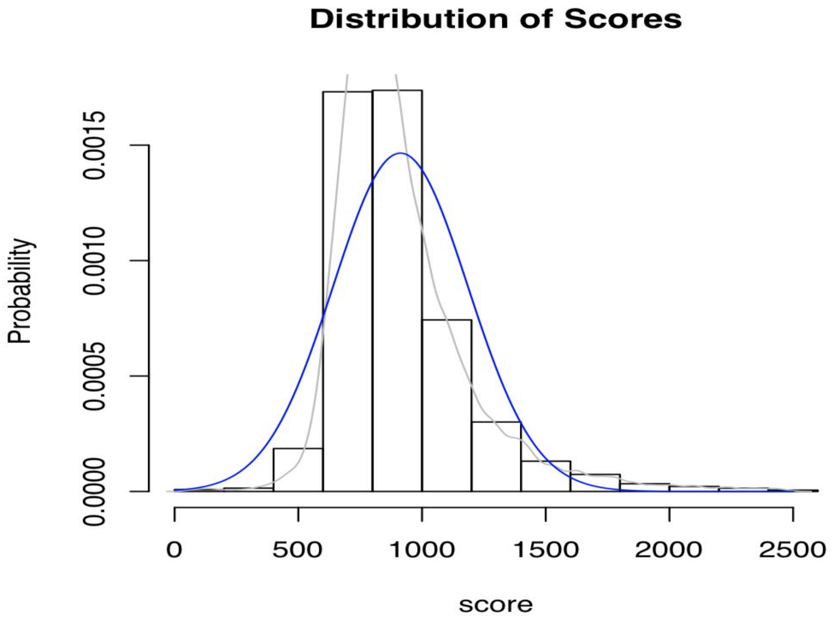


Figure 2B. Age-Related Differences in Performance (Median Reaction Time)

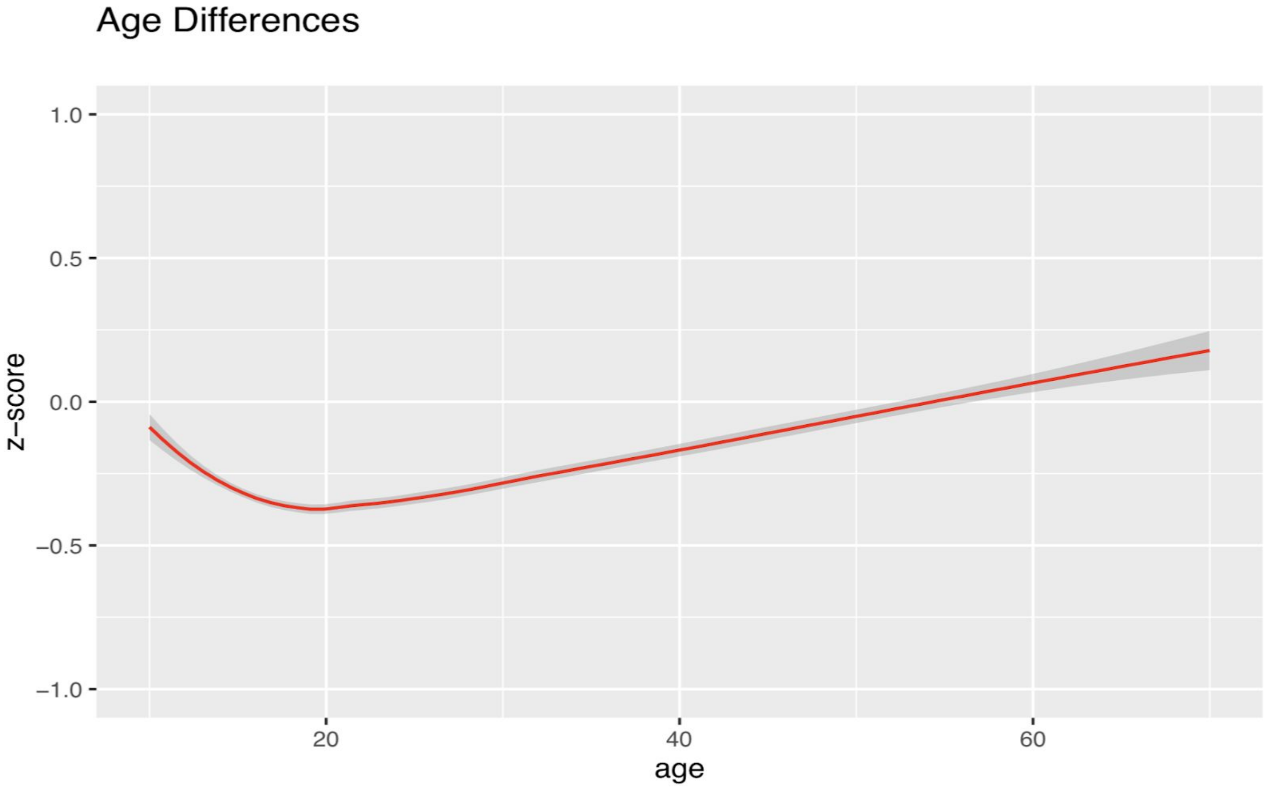


Figure 2C. Sex Differences in Performance (Median Reaction Time)

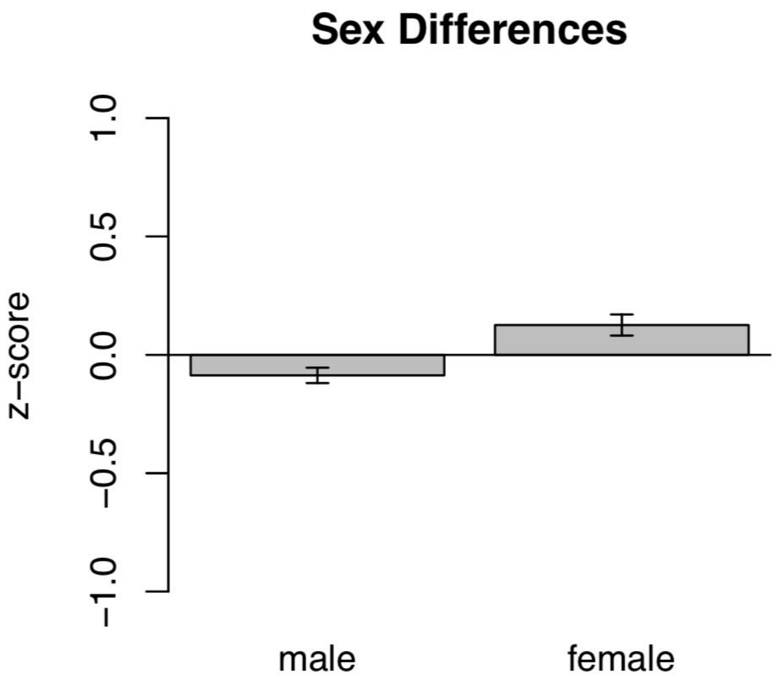


Figure 2D. Education-Related Differences in Performance (Median Reaction Time)

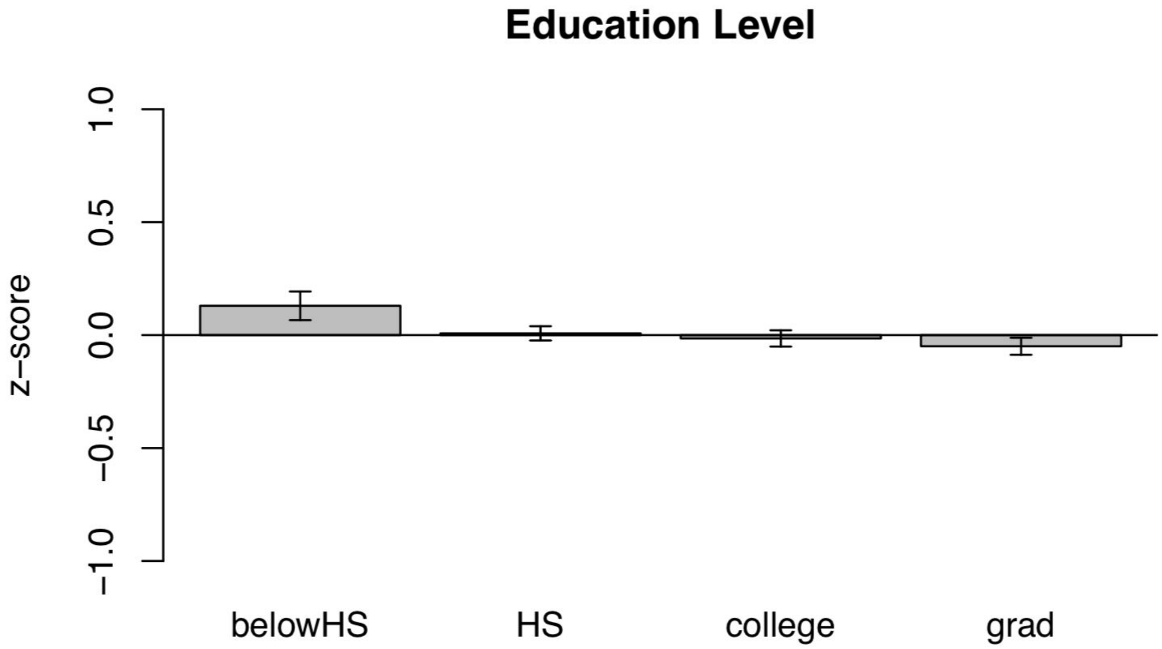


Figure 2E: Distribution of Scores (IES)

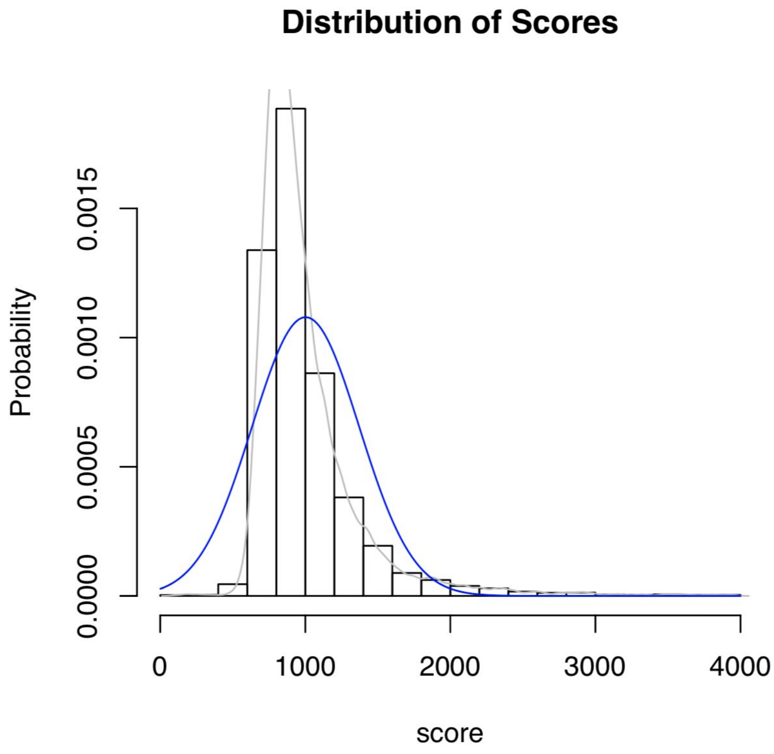


Figure 2F: Age-Related Differences in Performance (IES)

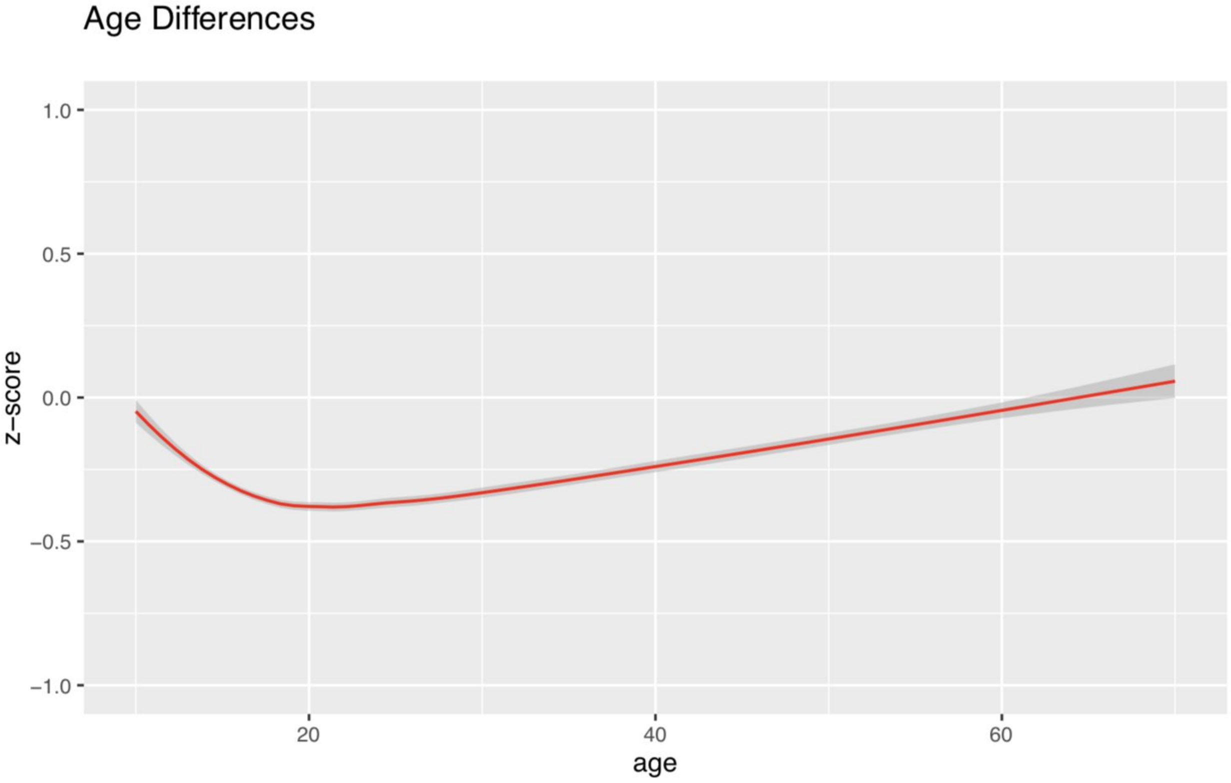


Figure 2G: Sex Differences in Performance (IES)

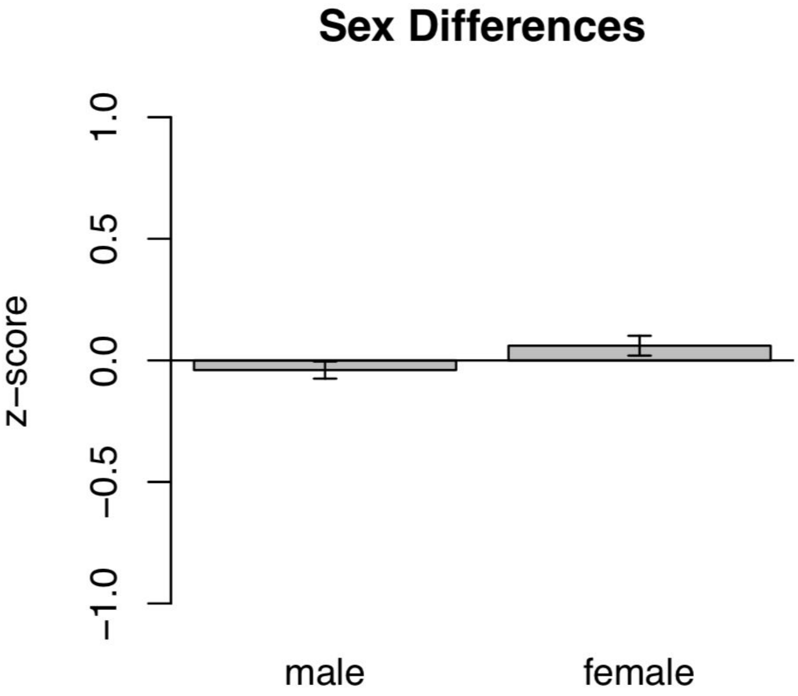
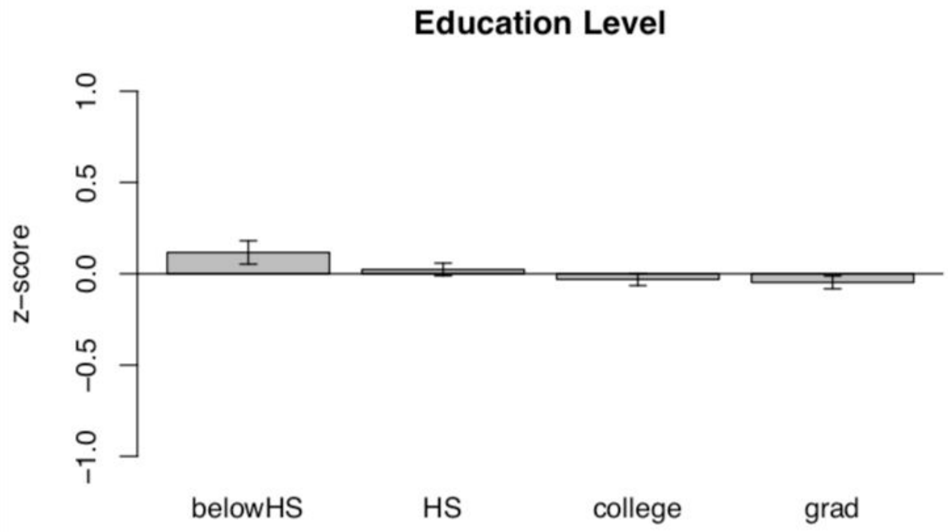
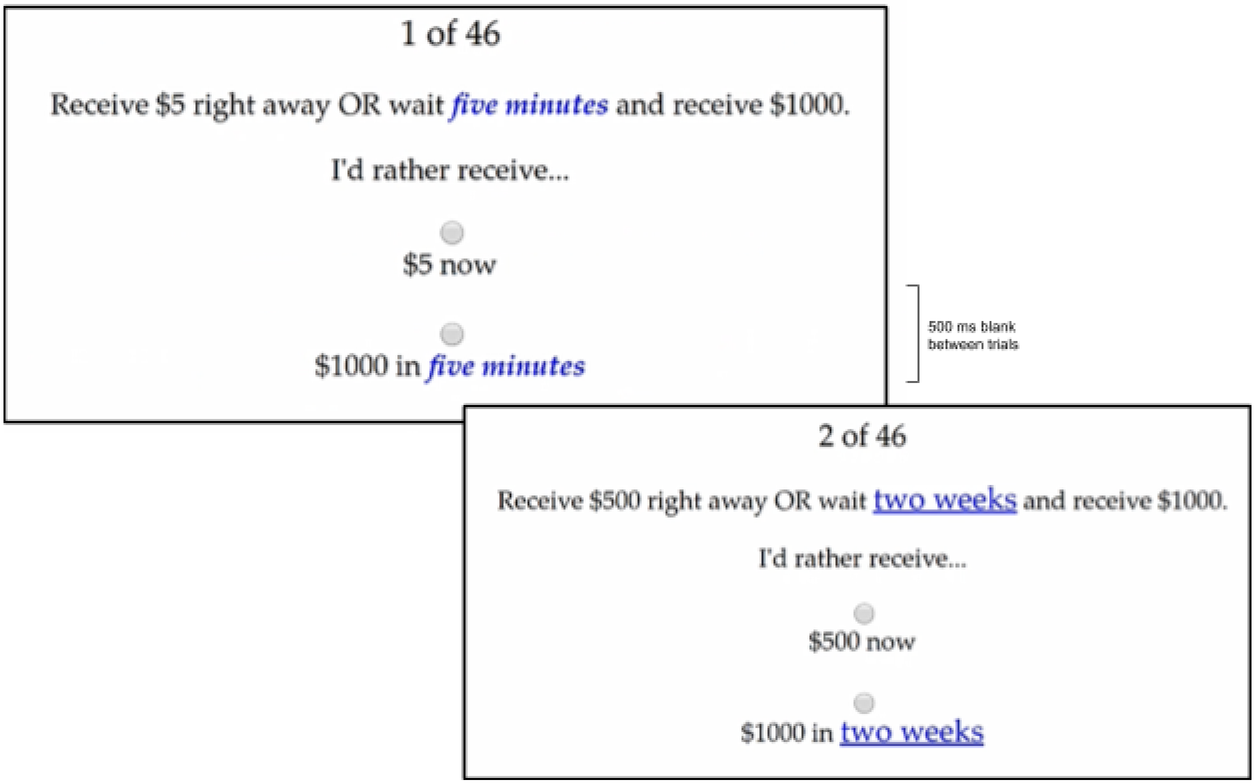


Figure 2H: Education-Related Differences in Performance (IES)



**TMB Delay Discounting**

Constructs Measured: temporal discounting, monetary decision-making
Duration: 5 minutes
Sample size for which normative data are available: 37,519
Demo Link: http://www.testmybrain.org/tests/delay_discounting/
Description of procedure: Compute how much you "value" your current self as compared to your future self. The value is computed using hypothetical monetary trade-offs, e.g.: Would you rather receive $20 now, or $80 in one year?



1 of 46

Receive $5 right away OR wait *five minutes* and receive $1000.

I'd rather receive...

○ $5 now

○ $1000 in *five minutes*

500 ms blank between trials

2 of 46

Receive $500 right away OR wait two weeks and receive $1000.

I'd rather receive...

○ $500 now

○ $1000 in two weeks

This is an adaptive format of a standard delay discounting task for estimating individual differences in temporal discounting (Kirby & Marakovic, 1995). Advantages of the test are that it is short and can be administered across a range of mobile devices.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Delay discounting tasks are included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 1.**

### Current Applications

A form of this test is currently included in NIMH Aurora study and the 23andme cognitive testing platform.

### Psychometric Characteristics

The Delay Discounting test measures the degree to which participants value immediate rewards over delayed rewards. This tendency is measured most directly by the coefficient *k* (range = 0, 1), which describes the relationship between the value of a future reward and the value of an immediate reward for a given delay length. A higher *k* value indicates a greater degree of discounting (a greater preference for immediate rewards over delayed ones). To create a measure of temporal discounting that can be presented more clearly to participants, the *k* parameter can also be transformed into a score representing the number of months it takes for a given amount of money to lose 50% of its “value”, where lower scores indicate greater or more rapid temporal discounting (score = (1/*k*)/30.375).

Scores on this test are highly reliable. Based on a sub-sample of 380 participants enrolled through the Aurora project, the internal reliability of this test is 0.92.

Sociodemographic effects were estimated based on score for the 34847 participants for whom demographic data was available. This sample had a mean age of 27.5 and was 50.5% female. The score distribution is skewed towards lower scores, indicating a higher rate of discounting and thus a greater preference for immediate rewards (see Figure 3A). Note this distribution is very different from what was found in the Aurora sample, where the modal response is a preference for delayed rewards. Discounting behavior is variable over the life course, with younger participants having lower mean scores (indicating greater temporal discounting) than older participants (see Figure 3B). Female participants have higher mean scores (indicating reduced temporal discounting) than male participants (see Figure 3C). Score increases slightly with level of education, but this effect is extremely small (see Figure 3D).

Our data show no evidence of practice effects in this test. First-time participants have a mean score of 21.1, compared to a mean score of 21.0 for repeat participants.

### Validation

The Delay Discounting test correlates modestly with several measures of risk-taking and impulsiveness, behaviors related to time preference. After adjustment for age, scores on this test correlate with scores on the Barratt Impulsiveness scale (r = -0.10, N = 6740, 95% CIs = [-0.12, -0.076]).

### Appropriateness for Field Test Use

*Device Effects.* The Delay Discounting test is essentially a questionnaire and so is easy to administer across a wide variety of digital devices. Our data does show differences in score between participants who used mobile phones to take the Delay Discounting test and those who used tablets, laptops, or desktop computers (iPhone mean = 23.9, SD = 32.1, N = 233; iPad mean = 24.1, SD = 49.5, N = 293; Macintosh laptop/desktop mean = 24.1, SD = 46.5, N = 1006), but these differences are very likely due to demographic differences between users of different devices. For instance, participants who used iPhones to take the test have a mean age of 31.4, while Macintosh users have a mean age of 35.7 years and iPad users have a mean age of 40.0. Because scores on this test do not depend on any time-based measures, there is little reason to expect that device differences would have a direct effect on performance.

*Participant Burden.* This task is well-tolerated by participants, but is not as engaging as other tests. The mean participant rating for batteries containing this test is 3.54 out of 5, compared to a site-wide mean of 3.7. 86.7% of participants who begin this test complete it, which is slightly higher than the site average (81%).

**Further Development**

The current version is likely unnecessarily long.  Given the very high reliability of the test, it would be possible to shorten it substantially (e.g. by 75%) while still maintaining acceptable reliability.

Figure 3A. Distribution of Scores

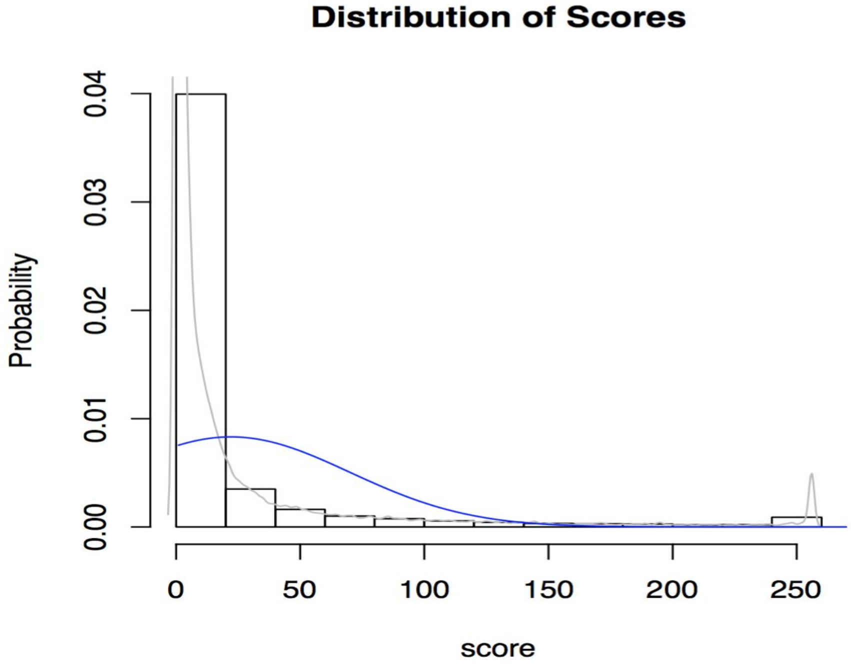


Figure 3B. Age-Related Differences in Performance

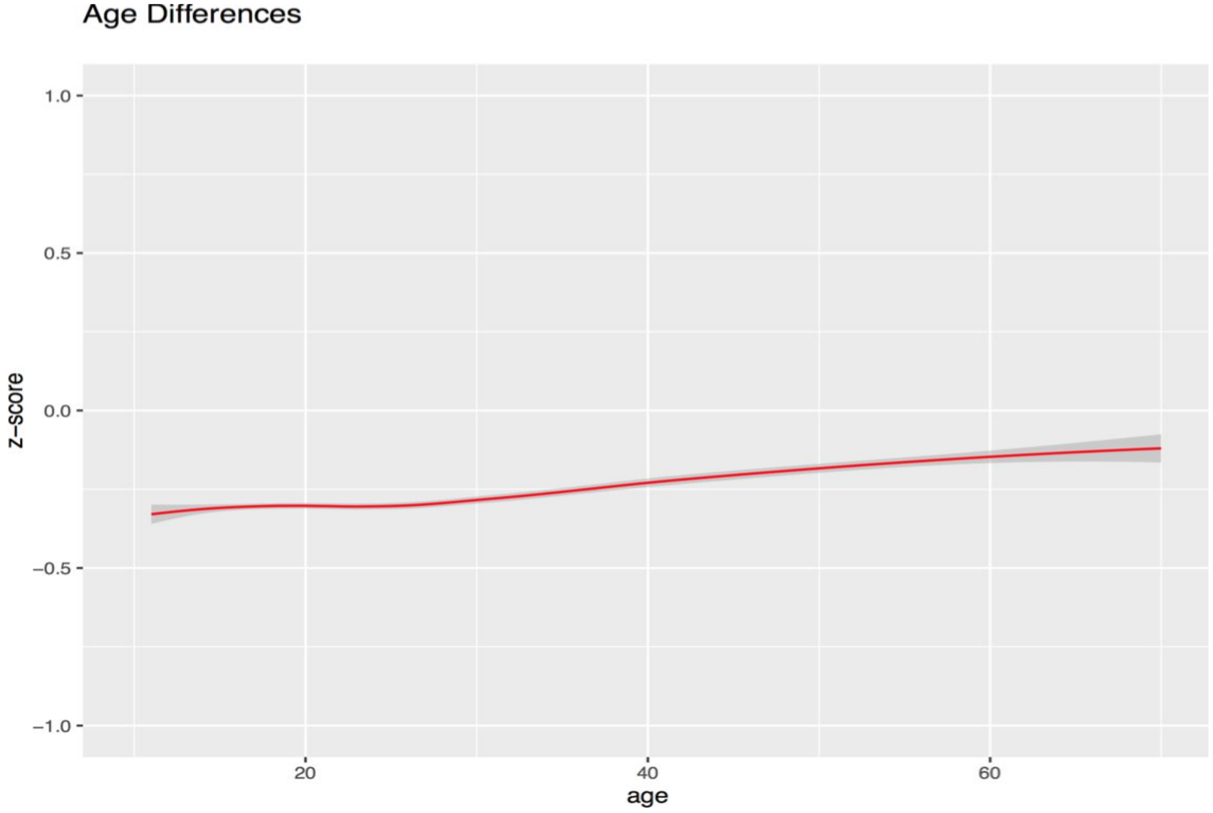


Figure 3C. Sex Differences in Performance

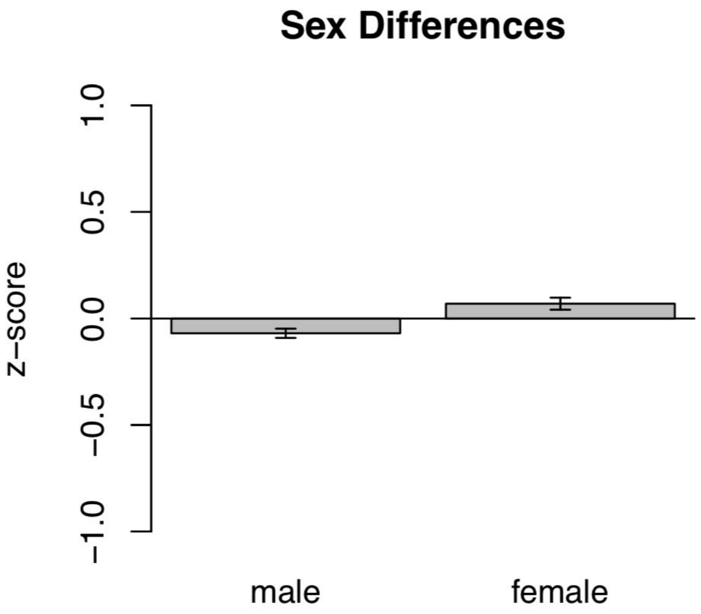
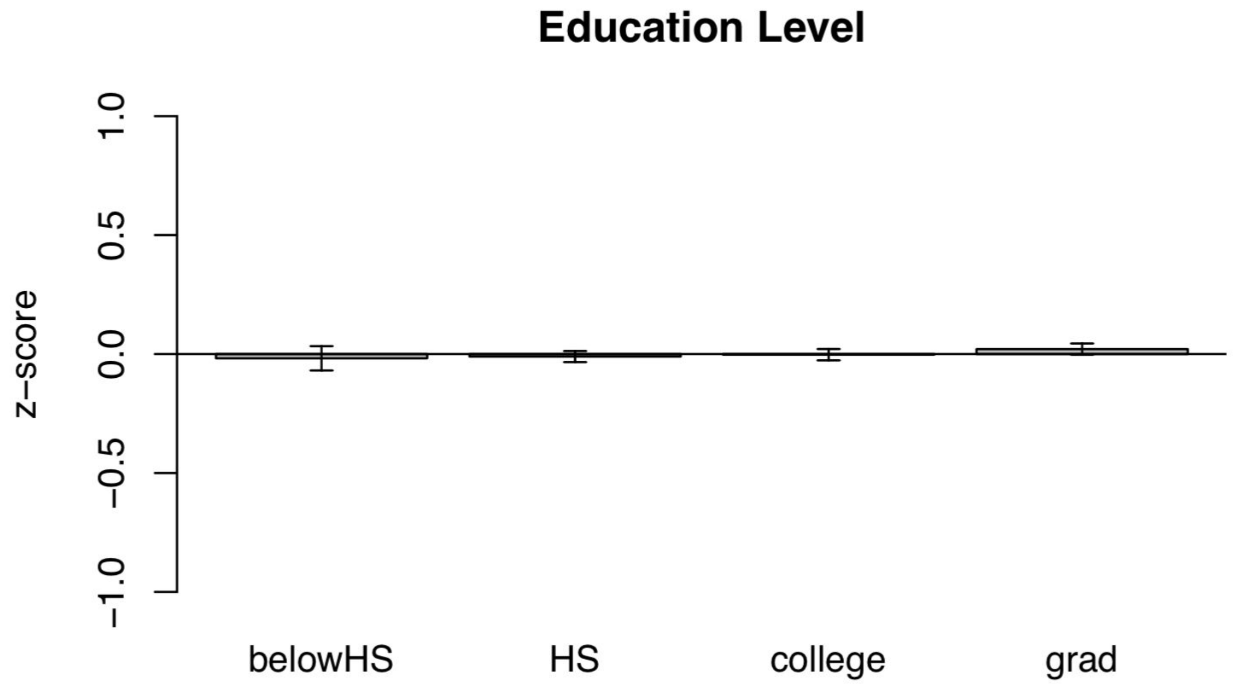
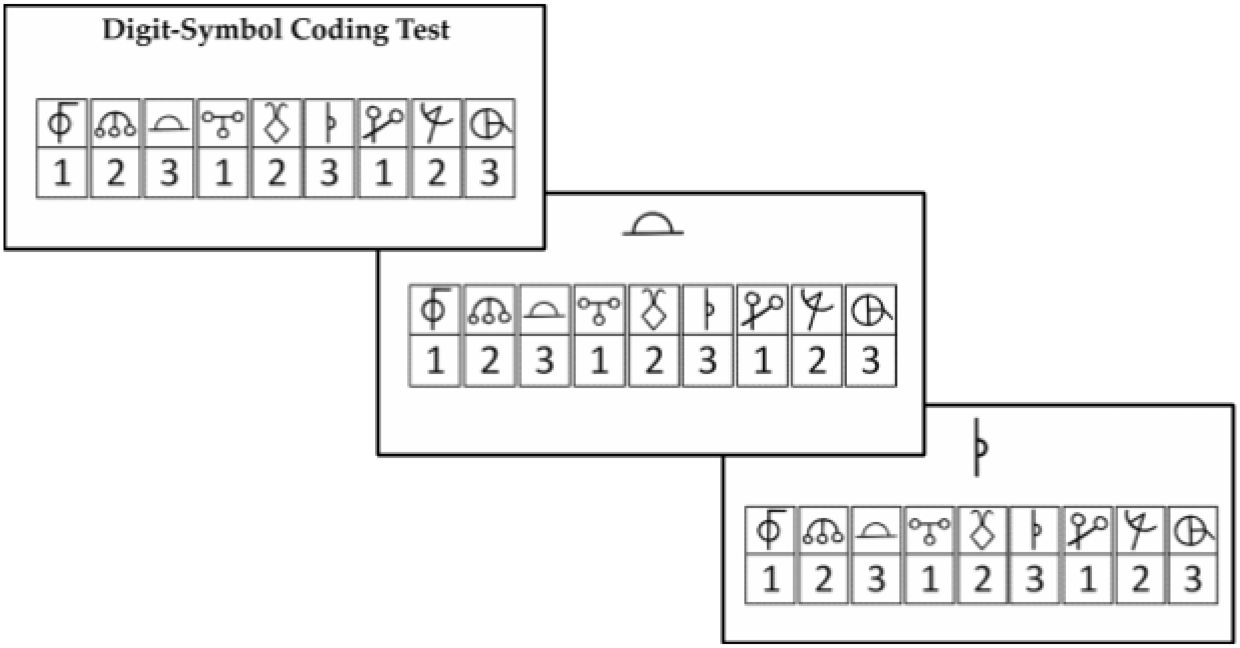


Figure 3D: Education-Related Differences in Performance



### TMB Digit Symbol Matching Test

Constructs Measured: processing speed, visual short term memory
Duration: 2.9 minutes
Sample size for which normative data are available: 45,295
Demo Link: http://www.testmybrain.org/tests/DigSymbCoding/DSC.html
Description of procedure: Using a symbol-number key shown on screen, match as many symbols and numbers as possible in 90 seconds.



This test is based on a well-validated and widely used measures of processing speed (e.g. WAIS digit symbol coding or digit symbol substitution tests) of a comparable format that have been used in clinical neuropsychology for decades (Joy, Kaplan & Fein, 2004). Advantages of the task are that it is very short, can be administered quickly and easily on a mobile device, and performance can be interpreted with respect to a large body of existing literature. Drawbacks are an inconsistent number of trials per individual, potentially complicating time series and standard psychometric analyses.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Although processing speed tests are not included in the RDoC Council Workgroup Report on Behavioral Assessments, we consider them important baseline measures for interpretation of other task data, so this task is designated **PRIORITY 1.**

### Current Applications

The TMB Digit Symbol Matching test is currently included in several major initiatives, including as part of the NIMH Aurora study, the Broad Neuropsychiatric Phenotyping Initiative, the NIDDK Brain Health Index, and the 23andme cognitive testing platform. Translation of the test into standard Chinese and Spanish is currently being funded by the Broad Institute.

### Psychometric Characteristics

The main outcome measure for this test is number of trials correctly completed in 90 seconds, which is proportionate to mean response time. Scores on this time are very reliable, with internal reliability (split-half) was 0.93 and test-retest reliability was 0.72 (calculated using data from the 1026 participants enrolled through the Aurora project, who took the test on multiple occasions).

Sociodemographic effects were estimated based on the pool of 40,977 participants for whom demographic data was availability. This sample had a mean age of 30.06 and was 45.70% female. The distribution of scores is normal (see Figure 4A). Score on this test is variable over the life course, with scores increasing (indicating faster reaction time) throughout adolescence and young adulthood, plateauing from approximately age 20 to age 30, and decreasing from age 30 into older adulthood (see Figure 4B). There are no significant differences in score between male and female participants (see Figure 4C). Scores increase somewhat with education, though this effect is not apparent in the most educated groups (see Figure 4D).

Our data show minimal practice effects for this test. Participants taking the test for the first time had a mean score of 48.22, while participants repeating the test had a mean score of 50.41 (Cohen's d = 0.15).

### Validation

The TMB Digit Symbol Matching task correlates with other measures of cognitive processing speed. It is moderately correlated with response speed in a simple reaction time test ($r = 0.32$, $n = 21023$) and a choice reaction time test that also requires participants to act quickly based on visual input ($r = 0.39$, $n = 12441$). It also shows moderate to high correlation with more complex tasks loaded on cognitive processing speed and visual perception, such as the TMB Flicker Change Detection task ($r = 0.48$, $n = 2641$), a letter and number trail-making task ($r = 0.48$, $n = 7145$), and the TMB Flanker task ($r = 0.35$, $n = 688$). The test is minimally associated with tasks that measure general cognitive ability but load minimally on short term memory and processing speed, such as vocabulary ($r = 0.03$, $n = 5248$). Performance on this task is correlated with depression symptoms, as measured by the Beck Depression Inventory ($r = 0.13$, $n = 294$, $p < 0.05$).

### Appropriateness for Field Test Use

This task, being brief and well-tolerated by participants, is well-suited to field test use. To ensure that participants understand the task before they begin, the test includes three practice trials before the test trials begin.

*Device Effects:* The Digit Symbol Matching test is easy to administer across a wide variety of device types. However, since this test measures cognitive processing speed using reaction time, differences in device performance (such as device latency in registering input) are likely to impact measured scores. Our data showed that participants using desktop or laptop computers had slightly higher scores (and thus lower reaction times) than those using mobile devices (iPhone mean = 46.73, SD = 9.09, N = 4615; iPad mean = 45.95, SD = 9.12, N = 2907; Macintosh desktop/laptop mean = 49.85, SD = 11.07, N = 10313). Thus, it appears that device latency may play a role in participant scores, with a Cohen's d = 0.4 differences between scores on Macintosh and iPad scores. Part of this difference may be due to sociodemographic differences, but device related variability will certainly also play a role.

*Participant Burden:* Digit Symbol Matching poses a low burden to participants and is generally well-tolerated. The average participant rating for batteries containing this task was 4.0 out of 5, compared to an average of 3.7 for all batteries hosted on Test My Brain. 93% of participants who began this test completed it.

**Further Development**

This test can be readily modified for ecological momentary assessment designs, and reliable scores measuring cognitive processing speed can be obtained in as little as 30 seconds. It is possible that differences between devices used to take this test may affect the measurement of scores, so when using this test it would be necessary to control for the devices used by participants. Otherwise, this test is ready for field test use.

Figure 4A. Distribution of Scores

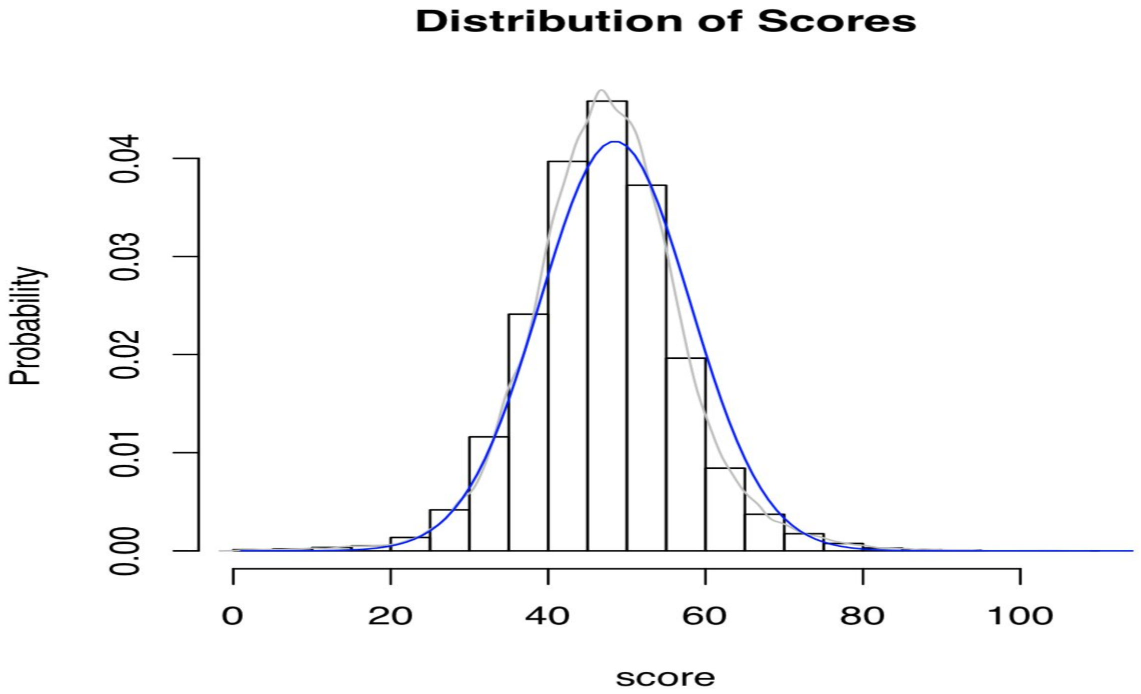


Figure 4B. Age-Related Differences in Performance

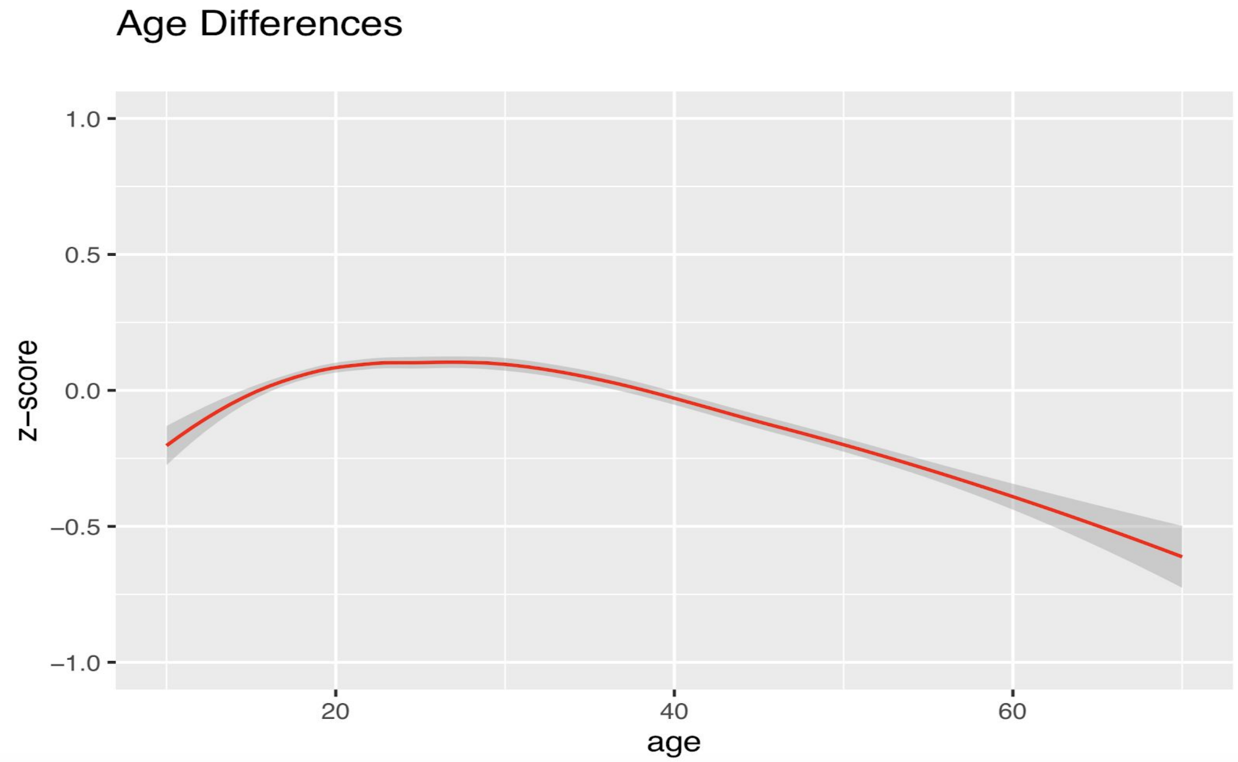


Figure 4C. Sex Differences in Performance

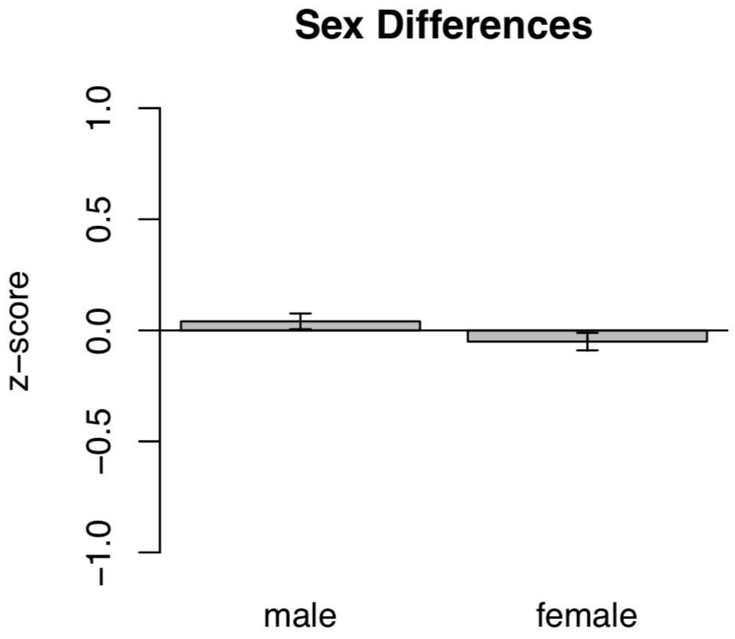
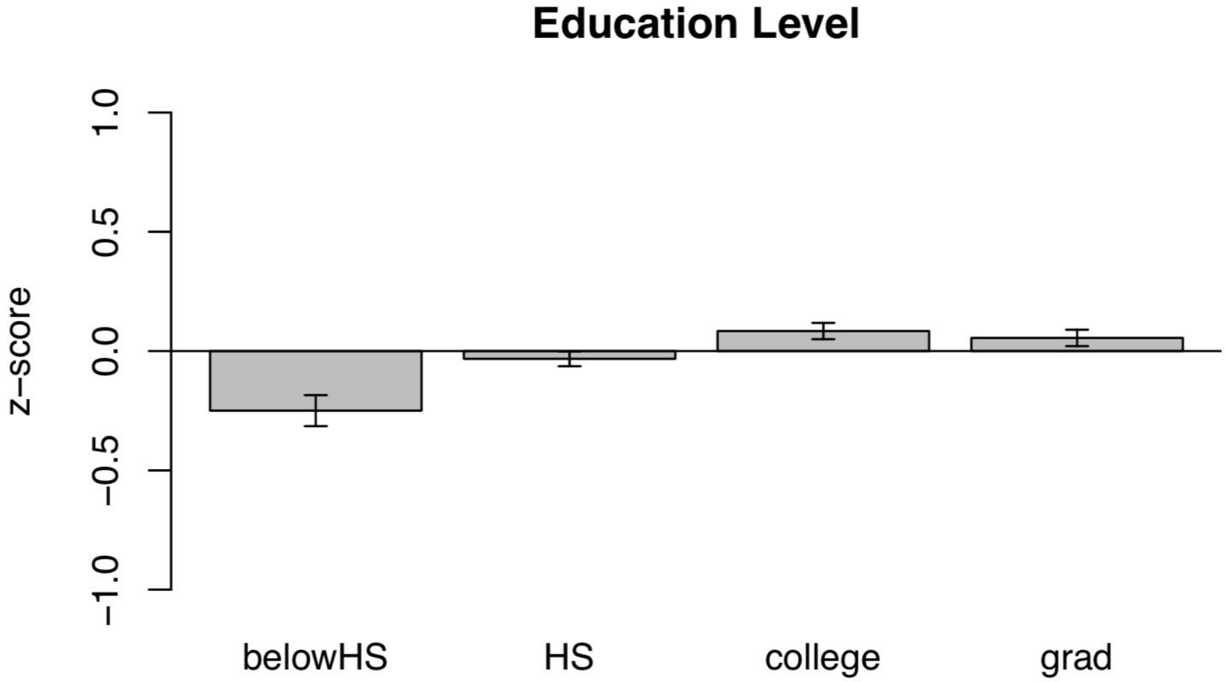


Figure 4D. Education-Related Differences in Performance



### TMB / Dillon Emotional Word Memory

Domains and *Constructs* measured:

1. Cognitive Systems: *declarative memory*
2. Social Processes: *perception and understanding of self* (subconstruct: self-knowledge)
3. Negative and Positive Valence Systems: Emotionally valenced negative and positive words are used.

This task is designed to comprehensively assess encoding, free recall, recognition memory, and source memory.

### Task

At **encoding**, the subject completes 100 trials on which a negative or positive word is presented; the word types differ significantly in valence but not on any other property (e.g., arousal, length, iter-item associativity). For each word, the subject must make one of two yes/no judgments: “Does this describe you?” or “Is this word positive?” At **recall**, the subject is asked to type as many words as he or she can remember. After recall, all the “old” words (i.e., from encoding) are presented again, intermixed with an equal number of matched “new” lures. During this **recognition memory** test, the subject must indicate which words are old vs. new. When a word is endorsed as “old”, the subject is prompted to indicate which judgment it was paired with at encoding—“describes?” or “positive?” This **source memory** test depends on the participant’s ability to retrieve contextual details from encoding. This battery thus assesses self-referential processing (“describes?” judgment), the ability to accurately assess emotional valence (“positive?” judgment), self-generated retrieval (free recall), plus recollection and familiarity (recognition and source memory).

This task is informed by decades of research on memory and all the tasks involved are well-established (Burt, Zembar, & Niederehe, 1995). The mobile-friendly version, however, was assembled recently and so we only have data from two samples. Nonetheless, the sample sizes are reasonably large by the standards of behavioral research (sample 1, *n* = 90; sample 2, *n* = 73). We have focused on establishing the basic patterns of the results, described next.

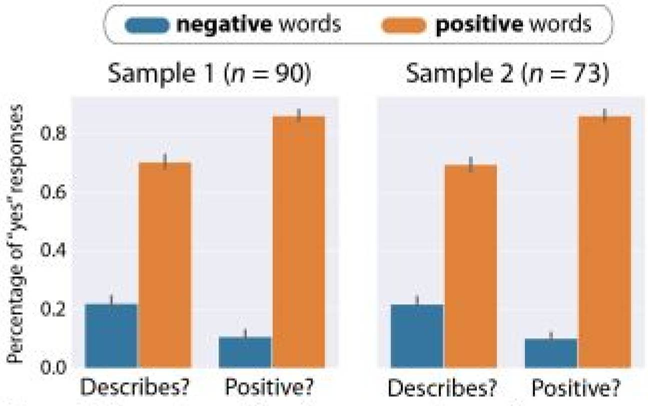


*Figure 1*. Percentage of ‘yes’ responses at encoding.

### Results

***Encoding***. The encoding data are shown in Figure 1. As can be seen, subjects endorse many more positive than negative words as self-descriptive, and their emotion judgments are mostly consistent with the normative data used to select the stimuli (i.e., about 90% of the positive words are judged to be positive). These two effects are very robust (*p*s < 0.001) in both samples.

***Free recall***. The free recall data are shown in Figure 2, which uses three rows to highlight a strength of our approach: by including two encoding tasks and sorting the data by encoding responses, we can drill down on the unique contributions made by emotion and self-referential processing to memory. The **top row** shows a significant recall advantage for positive vs. negative words in both samples. This is the typical pattern in healthy samples and is often assumed (e.g., in Dillon, 2015) to reflect a beneficial effect of positive emotional responses on encoding. The **middle row** demonstrates that things are actually more complicated. In Sample 1 at least, the effect of emotion on memory is restricted to words from the “describes?” task; memory for negative vs. positive words is equivalent for words from the “positive?” task. This demonstrates that the “emotion” effect on memory in this sample is contingent on self-referential processing (and so may not really be an emotion effect at all). Finally, the **bottom row** shows that this more nuanced conclusion is actually contingent on another factor—namely, the response the subjects made during encoding. For example, notice that for words from the “describes?” task, the positive memory advantage only holds for words endorsed as self-descriptive (“yes” responses). By contrast, for words from the “describes?” task that elicited a “no” response, there is a recall advantage for negative vs. positive words in both samples. In other words, the effect of valence on memory for words from the “describes?” task flips depending on the encoding response. In short, these data show that this task can be used to disentangle the impact of emotion, self-reference, and encoding response on recall. It is clear from many studies that, relative to healthy controls, adults with depression (and many other forms of psychopathology) have a substantially more negative self-image and are likely to show biased emotional responses at encoding (exaggerated for negative, blunted for positive). This paradigm provides a method for determining precisely how those psychological phenomena affect subsequent memory.

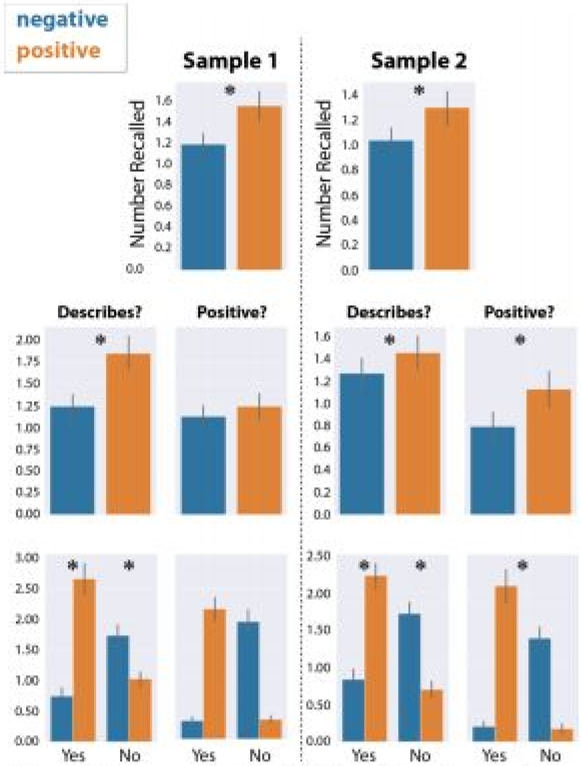


*Figure 2*. Free recall data. Effects of *Emotion* (top), *Emotion x Task* (middle), and *Emotion x Task x Response* (bottom). *$p < 0.05$.

***Recognition and Source memory***. Figure 3 shows recognition memory accuracy, assessed with the signal detection measure *d*' (top), and source memory accuracy (bottom), assessed by hit rate. Again, source memory is assessed by asking the subject to indicate which encoding judgment (“describes?” or “positive?”) was made for any word endorsed as old. The figure shows that recognition is significantly better for positive vs. negative words in Sample 2; the effect in Sample 1 is marginal (*p* = 0.056). By contrast, source accuracy is better for positive vs. negative words in both samples. This is interesting because recognition memory is supported by both recollection and familiarity, whereas source memory depends almost solely on recollection (because it depends on the recovery of contextual details from encoding). To the extent that recollection preferentially recruits the hippocampus, these results thus suggest that emotion may have a reliable effect on hippocampal function in this task.

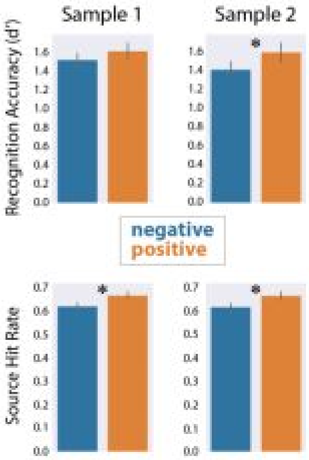


*Figure 3*. Recognition (top) and source (bottom) memory accuracy.

**Participant burden and psychometrics**

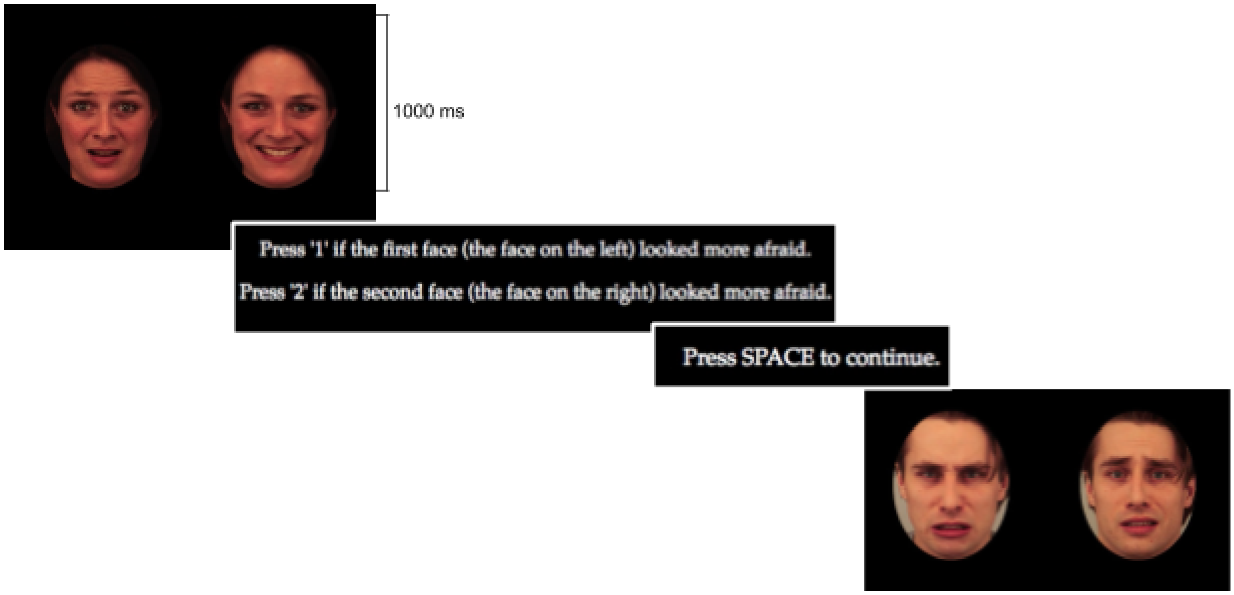
The primary weakness of this task is its length—it took participants in Sample 1 about 17 minutes to complete (encoding: 4 min; recall: 3 min; recognition:  10 min). Consequently, the attrition rates were quite high: 56% in Sample 1, 58% in Sample 2. Subjects who completed the task rated it as tolerable (3.51 and 3.52 of 5.00 stars in Samples 1 and 2), but it clearly needs to be shortened. This should not be problematic, as removing a third or even half the trials would still leave as many as 50 encoding trials, which should be sufficient to see the effects documented above. Another clear need is an assessment of the psychometrics of the tasks, which is somewhat more complex in this domain than in tasks in which the stimuli do not vary from trial-to-trial.

**Conclusion**

Episodic memory is central to our lives. We rely on it daily in mundane circumstances (“Where did I park my car?”), but it is also fundamental to the stories we tell about ourselves (“Did I have a happy childhood?”) As this last example shows, episodic memory is shaped by—and shapes—our emotional experiences and our sense of self. Moreover, all of these things—memory, emotion, self-concept—are affected by psychopathology (Matt, Vásquez, & Campbell, 1992). This task merits consideration for inclusion in an RDoC test battery because it offers a comprehensive, easily quantified assessment of all of these import psychological processes. Going forward, we will shorten the task to reduce participant burden, and we will assess the psychometrics of each aspect of the task (encoding, recall, recognition, and source memory) to determine which measures are more versus less reliable.

### TMB Fear Sensitivity

Constructs Measured: potential threat, social communication/reception of facial communication, understanding mental states
Duration: 2.4 minutes
Sample size for which normative data are available: 13,438
Demo Link: http://www.testmybrain.org/tests/emotion_comparison/fear.html
Description of procedure: Judge which of two faces is more fearful.



Press '1' if the first face (the face on the left) looked more afraid.
Press '2' if the second face (the face on the right) looked more afraid.

Press SPACE to continue.

This task assesses sensitivity to differences in fear intensity, independent of response bias and differences in emotion identification or categorization (Rutter et al., 2019). Advantages of the task is it allows issues related to categorization, verbalization, response bias to be dissociated from sensitivity to specific face emotions. It is also short and easy to administer across a range of mobile device types. Disadvantages are that the task is not yet validated with respect to clinical conditions or psychopathology and is considered burdensome by participants despite its relatively short length.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Emotion sensitivity or emotion comparison tests are not included in the RDoC Council Workgroup Report on Behavioral Assessments, so we consider this test **PRIORITY 2.**

### Current Applications

All three TMB Emotion Sensitivity tests are being used and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

The primary outcome measure from this test is accuracy, based on proportion or number correct out of 56 trials. This score reflects the participant's ability to detect differences in happiness between faces. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures is less clear.

This test shows good reliability, especially given its length. Internal reliability (split-half) was 0.80, as calculated from the scores a sample of 5000 participants who completed the test on TestMyBrain.

Sociodemographic effects were estimated based on the scores of the 12,072 participants for whom demographic data was available. This population has a mean age of 27.23 and is 59.67% female. The distribution of scores is relatively normal with some ceiling effects (see figure 5A). Performance is variable across the lifespan; scores increase during adolescence and plateau throughout adulthood before declining after age 50 (see figure 5B). Female participants show slightly higher performance than male participants (see figure 5C). Performance increases with level of education (controlled for age), though this effect is not consistent between the most highly educated participant groups (see figure 5D).

This test does not show evidence of practice effects. First-time participants have a mean score of 44.62, while repeat participants have a mean score of 42.72.

### Validation

Performance on this test is correlated with other tests of emotion perception. It shows moderate to high correlation with performance on analogous tests of perception of anger (r = 0.53, N = 12312, 95% CI [0.52, 0.54]) and happiness (r = 0.47, N = 11933, 95% CI [0.46, 0.49]). Scores do not correlate with anxiety scores as assessed by the GAD-7 (r = 0.015, N = 535, 95% CI [-0.070, 0.10]) or depression symptoms as assessed by the Beck Depression Inventory (rho = -0.077, N = 486, 95% CI [-0.16, 0.012]).

### Appropriateness for Field Test Use

Before beginning the test, each participant completes easy 2 practice trials, which include immediate feedback and are repeated if the participant answers incorrectly. Therefore, difficulties in understanding the task should not present a barrier to completion.

*Device Effects.* Participants who took this test using mobile devices showed slightly lower performance than those who used laptop or desktop computers (iPhone mean = 44.19, SD = 6.34, N = 1476; iPad mean = 43.98, SD = 5.91, N = 591; Macintosh laptop/desktop mean = 46.19, SD = 5.88, N = 1438). Device type may have an impact on performance on this test; for instance, although comparable scores between iPhone and iPad suggest that screen size does not explain this difference, but may instead be due to differences in demographics or environmental context.

*Participant Burden.* This test is considered burdensome by participants. The mean participant rating for batteries on TestMyBrain containing this test is 3.40, compared to a site-wide mean participant rating of 3.7. Completion rates are 67.3%, which compares unfavorably with other tests on TestMyBrain.org (81% completion average).

### Further Development

The current version of this test relies on faces taken from predominantly Caucasian face databases, so the major limitation of this test is its use in diverse cohorts.  Versions of the test that include multiracial faces are recommended for broader applications.

Figure 5A. Distribution of Scores

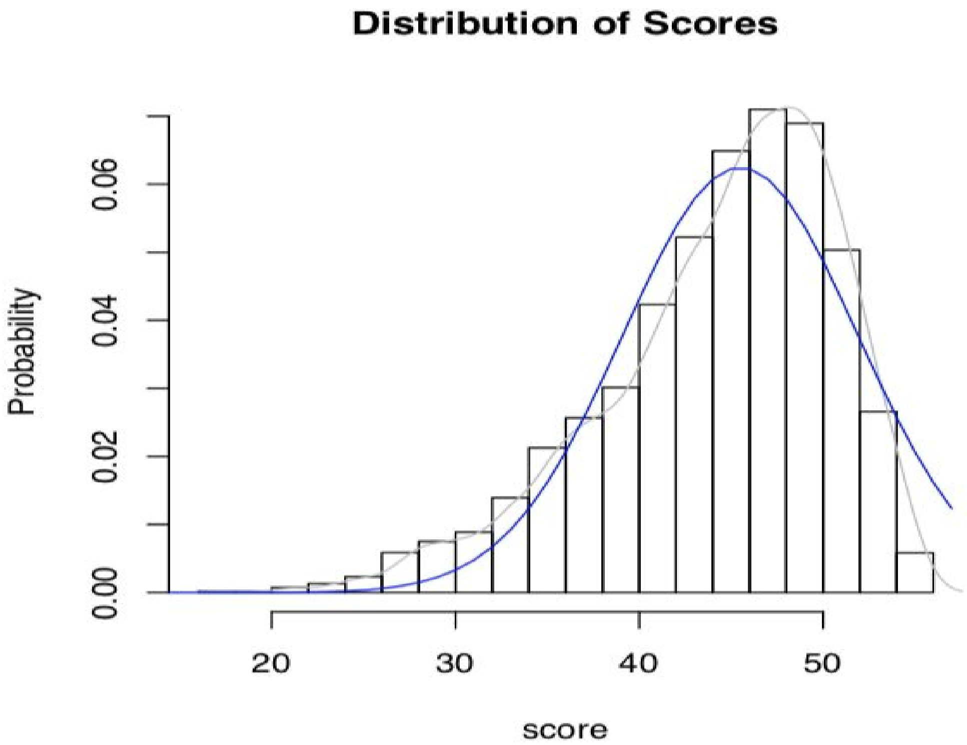


Figure 5B. Age-Related Differences in Performance

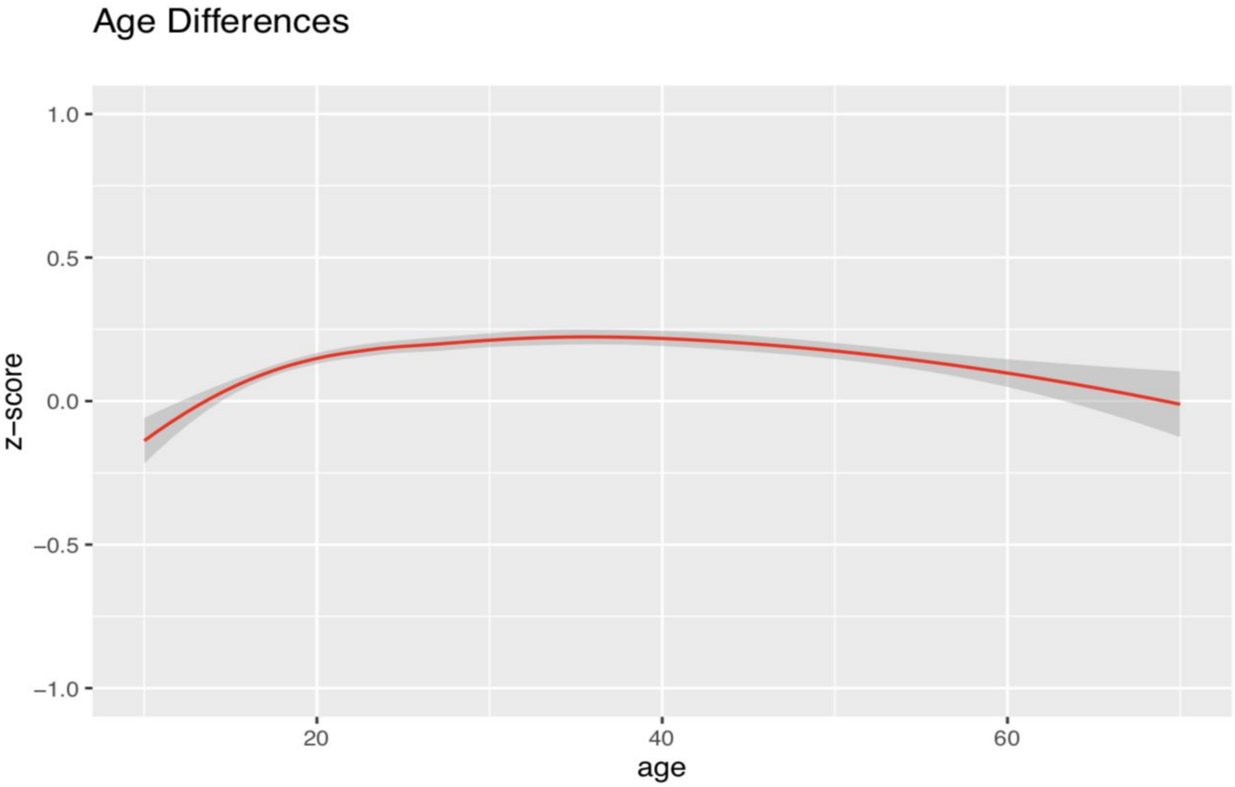


Figure 5C. Sex Differences in Performance

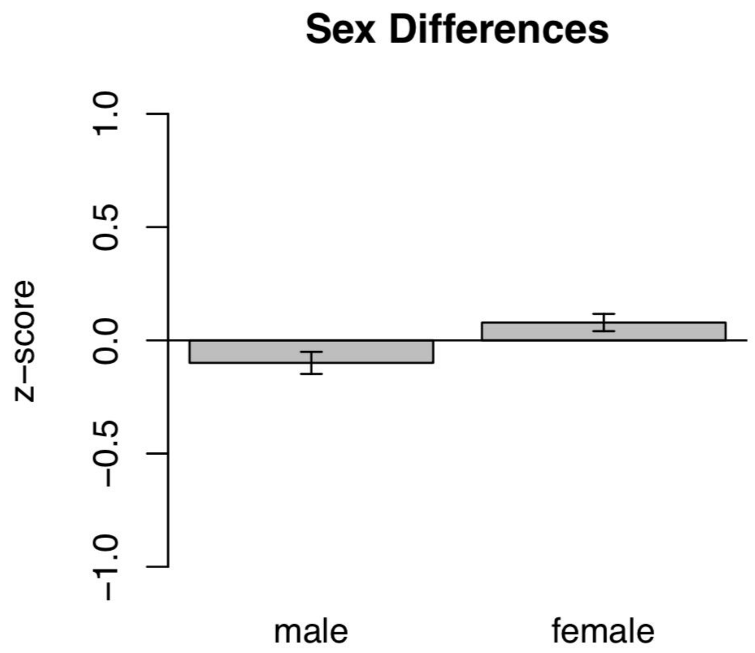
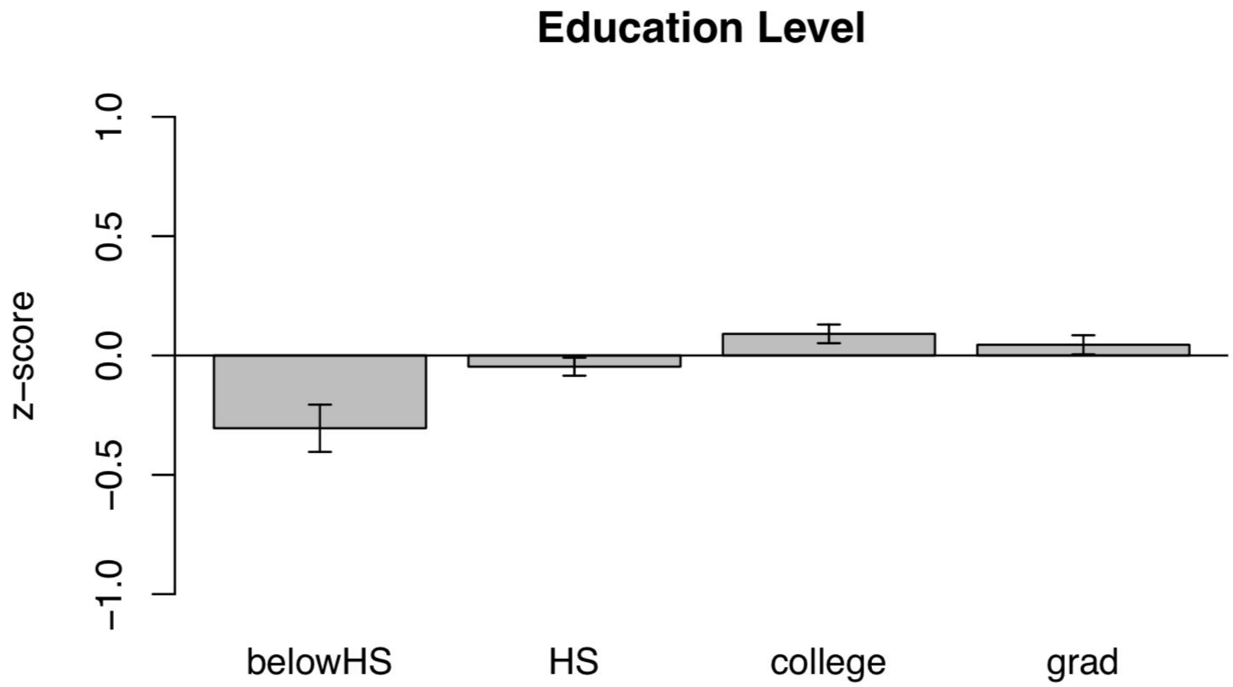
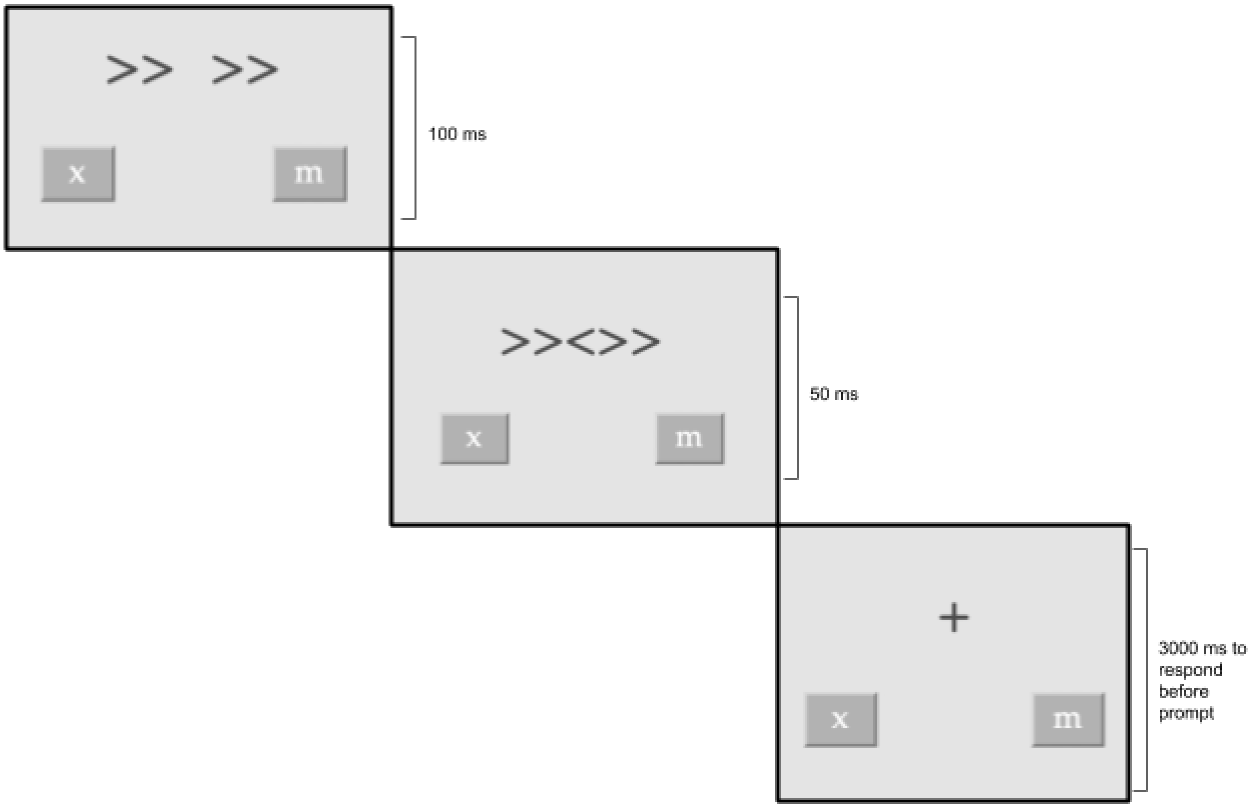


Figure 5D. Education-Related Differences in Performance



### Dillon/TMB Flanker Test

Constructs Measured: attention, cognitive control
Duration: 5.6 minutes
Sample size for which normative data are available: 10,885
Demo Link: http://www.testmybrain.org/tests/flanker/Flanker6.html
Description of procedure:Judge the direction of a central arrow flanked by four other arrows pointing in either the same direction (congruent) or a different direction (incongruent).



This is a standard Eriksen flanker task. Advantages are that the task is short and can be completed across a range of mobile devices. Disadvantages are that participants can find the task burdensome. The very brief display time of the central arrow increases the sensitivity of the task (producing accuracy effects related to flanker interference), but may be problematic for those with perceptual difficulties, and can be subject to stimulus presentation errors (e.g. appearing on screen too long) if there are competing processes on the user's computer.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Flanker tests are included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 1.**

## Current Application

This test is not included in any major initiatives, due to its relatively burdensome nature.

## Psychometric Characteristics

The Flanker task can yield multiple useful measures of cognitive processing speed, attention, and cognitive control (Dillon et al., 2015). Here, we focus on flanker interference effects on reaction time and accuracy, calculated as (1) RT conflict scores: the difference in mean reaction time between incongruent and congruent trials, as the main outcome measure, and (2) accuracy conflict scores: the difference in accuracy between incongruent and congruent trials.

Based on a sample of 2656 participants tested through TestMyBrain.org, RT conflict scores (mean difference between congruent and incongruent trials 70ms) showed modest reliability, with split-half reliability of 0.52. Accuracy conflict scores (mean difference between congruent and incongruent trials 11%) were more reliable, with split-half reliability of 0.77.

Sociodemographic effects were estimated based on RT conflict scores from the sample of 10,559 participants for whom demographic data is available. This participant group had a mean age of 30.93 and was 51.77% female. The distribution of scores was relatively normal for both RT and accuracy conflict scores with minimal restriction of range (see Figure 6A). Performance is variable across the life course, with age most strongly related to RT conflict scores (even controlling for accuracy), which decrease throughout adolescence and young adulthood before increasing steadily after approximately age 25 (see Figure 6B). Female participants have slightly higher mean RT conflict scores than male participants (see Figure 6C). RT conflict score decreases with increased education (see Figure 6D).

Accuracy conflict scores show a similar pattern of sociodemographic trends. These scores are also normally distributed (see Figure 6E) and variable across the life course, with scores decreasing until approximately age 30 and increasing after approximately age 50 (see Figure 6F). Female participants have slightly higher accuracy conflict scores than male participants (see Figure 6G). Like RT conflict scores, accuracy conflict scores decrease with increased education (see Figure 6H).

Neither RT conflict scores nor accuracy conflict scores show evidence of practice effects. First-time participants had a mean RT conflict score of 78.72 ms and a mean accuracy conflict score of 10.3%, while repeat participants had a mean RT conflict score of 78.50 ms and a mean accuracy conflict score of 10.1%.

## Validation

Based on a sample of 912, Flanker accuracy (but not RT) conflict scores were significantly associated with scores on the Beck Depression Inventory (BDI-II) ($B = 0.09$, $p < 0.01$). Accuracy conflict scores also show a small but significant negative correlation with morphed emotion identification ($r = -0.099$, $n = 2289$, 95% CI [-0.13, -0.049]).

## Appropriateness for Field Use

*Device Effects.* This test shows minor differences in reaction time conflict score based on the device used to take the test, with users of laptop and desktop computers having slightly lower conflict scores than users of mobile devices (iPhone mean = 94.66, SD = 71.20, N = 1182; iPad mean = 101.65, SD = 79.04, N = 724, Macintosh laptop/desktop mean = 86.63, SD = 79.40, N = 1840). Accuracy conflict scores show minimal device effects (iPhone mean  = 0.11, SD = 0.15, N = 482; iPad mean = 0.12, SD = 0.16, N = 312; Macintosh desktop/laptop mean = 0.10, SD = 0.15, N = 753). Since these are difference score, differences in device latency (the time it takes a device to register input) are unlikely to affect the measurement of scores. However, because the stimulus in this test is presented for a very brief time, problems impacting the display could have significant impact on the presentation of each trial and participant performance. Differences in screen size may similarly affect participants' ability to perceive the images presented in the course of the test.

*Participant Burden.* This test is considered somewhat burdensome by participants. The mean participant rating for batteries containing this test is 3.57 out of 5, compared to a site-wide mean participant rating of 3.7. 68.7% of participants who begin this test complete it, compared to 81% site-wide.

**Further Development**

The current version of this test went through several phases of development to find a version that was minimally reliable and burdensome.  In general, it is difficult to make tests of this type low burden as they require a relatively large number of trials to reach acceptable levels of reliability and - at the same time - are repetitive and not considered engaging by participants. Future versions might attempt to improve the engagement characteristics of the task, perhaps through other incentives.

Figure 6A. Distribution of Scores (RT Conflict Scores)

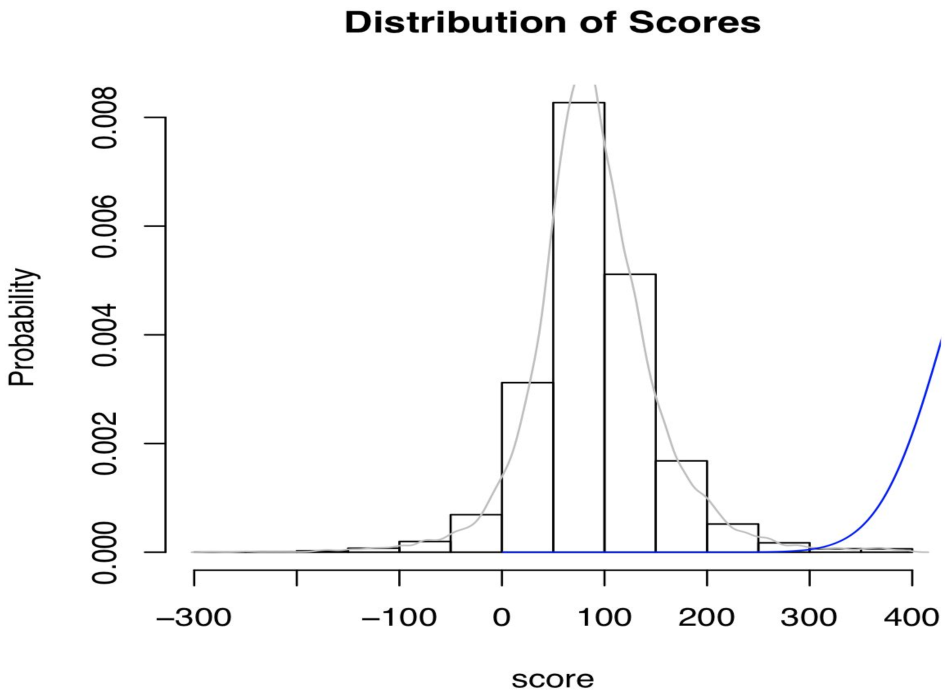


Figure 6B. Age-Related Differences in Performance (RT Conflict Scores)

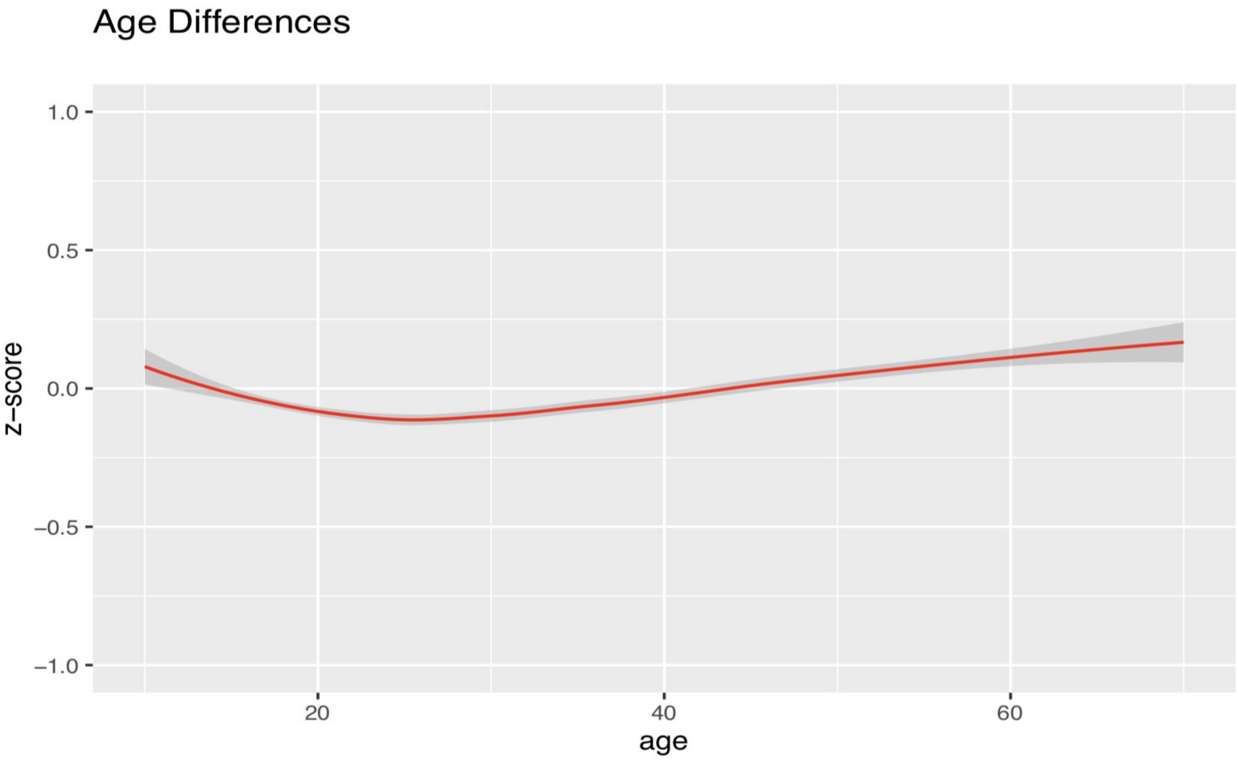


Figure 6C. Sex Differences in Performances (RT Conflict Scores)

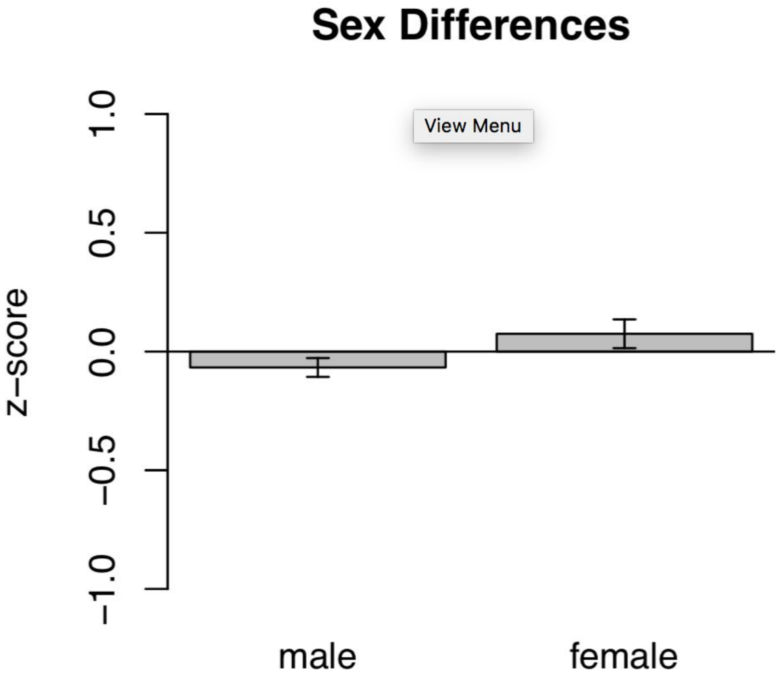


Figure 6D. Education-Related Differences in Performance (RT Conflict Scores)

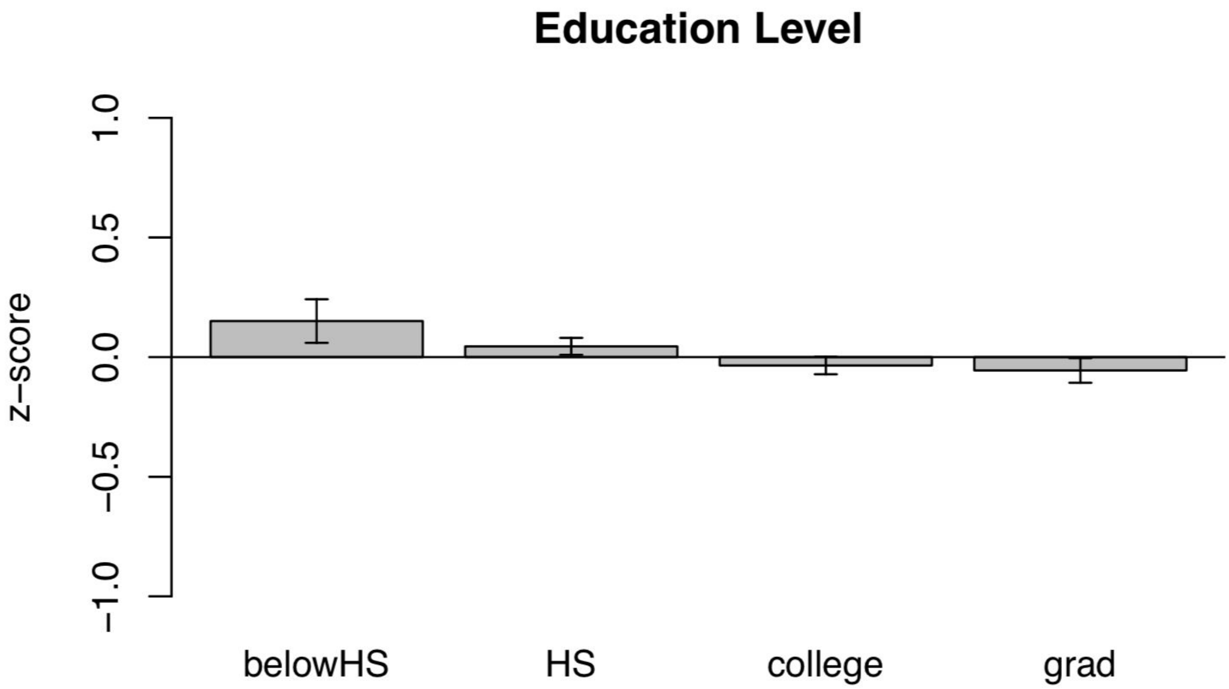


Figure 6E. Distribution of Scores (Accuracy Conflict Scores)

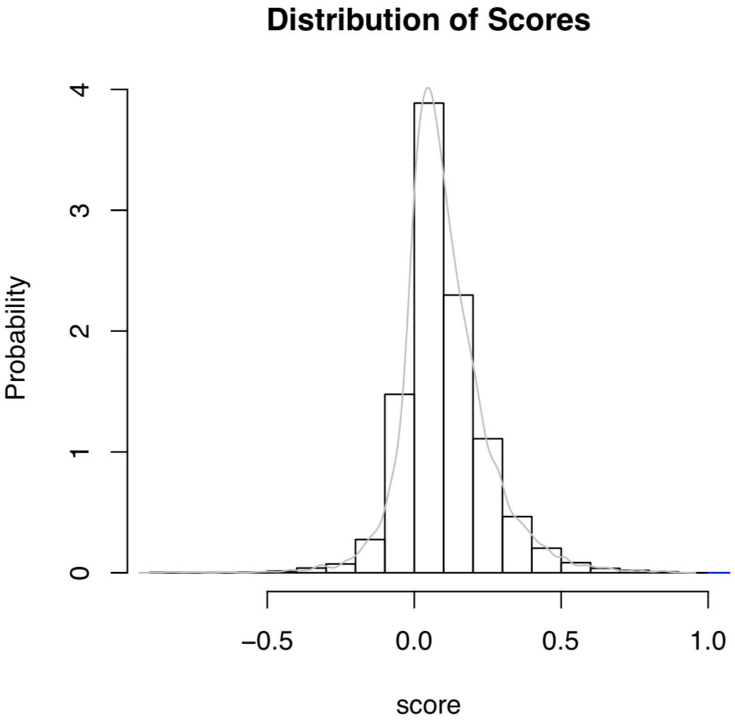


Figure 6F. Age-Related Differences in Performance (Accuracy Conflict Scores)

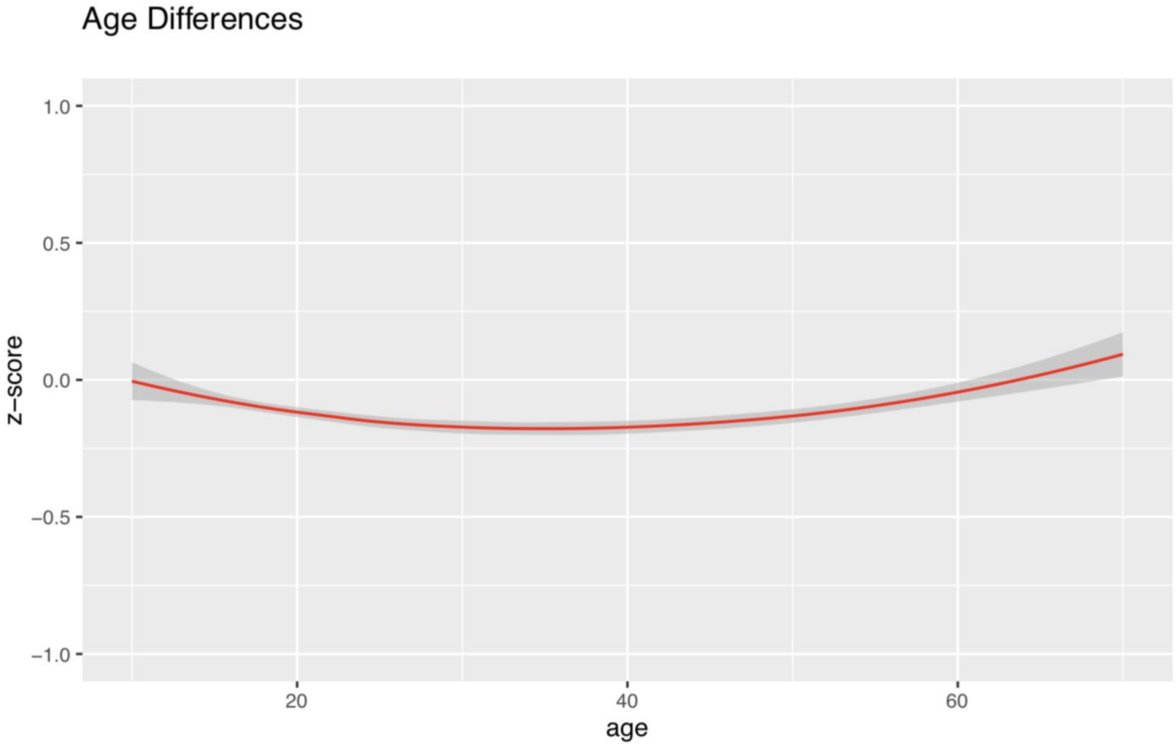


Figure 6G. Sex Differences in Performance (Accuracy Conflict Scores)

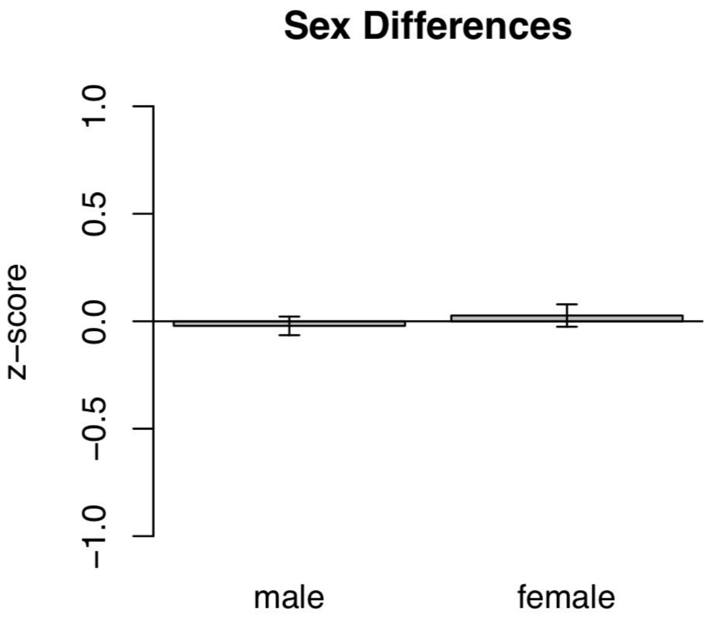
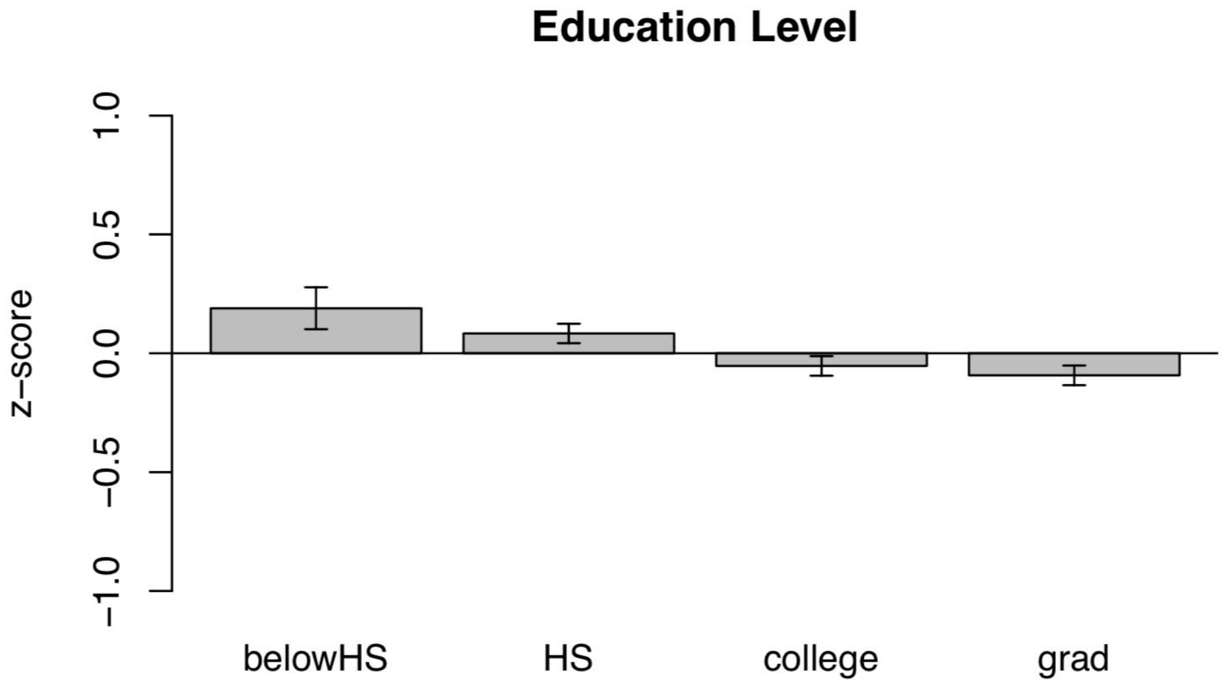


Figure 6H. Education-Related Differences in Performance (Accuracy Conflict Scores)



### TMB Flicker Change Detection Test

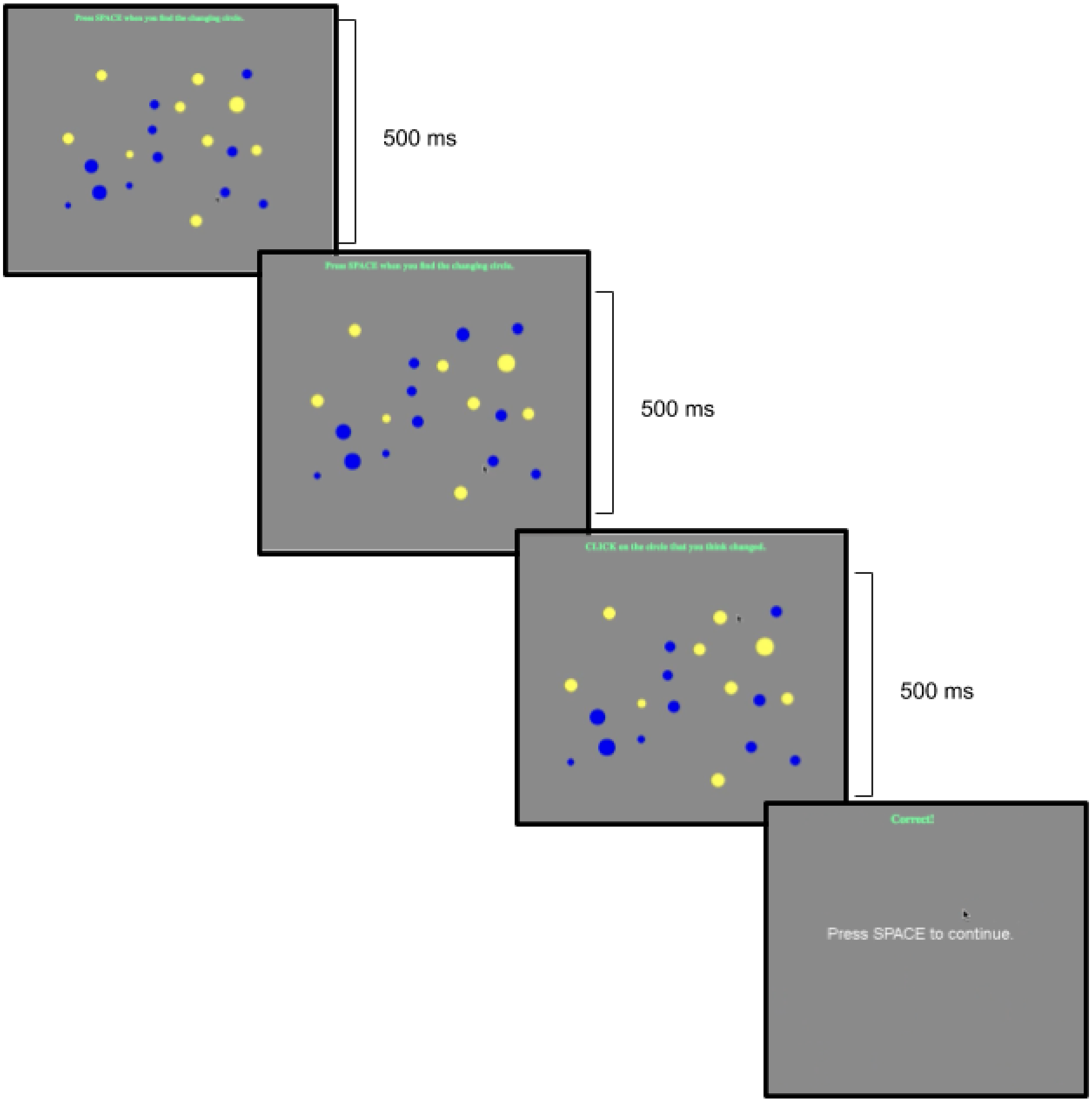
Constructs Measured: change detection, visual search

Duration: 5 minutes

Sample size for which normative data are available: 29,627

Demo Link: https://testmybrain.org/tests/flicker/flicker/flicker.html

Description of procedure: Given a set of flashing blue and yellow dots, find the dot that is changing color from blue to yellow.



Press SPACE when you find the changing circle.

Press SPACE when you find the changing circle.

CLICK on the circle that you think changed.

Correct!

Press SPACE to continue.

Images initially alternate every 500 milliseconds, with a 200 millisecond blank screen separating them. After 200 flips, a warning is displayed for 20000 milliseconds. If the participant presses space but does not respond after 15000 milliseconds, the same warning is displayed.

This is a visual search and change detection test adapted from the classic Resnick change detection test, but using a more precisely controlled stimulus set (Wilmer et al., 2012). Advantages of the task are that it is short, can be administered quickly and easily on a mobile device, and is considered enjoyable by participants. Disadvantages are a potential advantage to poor performance (as it makes the test shorter) and difficulties in interpretation due to the conflation of visual search with change detection. There are also may be potential limitations in stimulus delivery on a smaller screen.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Visual search and change detection tasks are included in the RDoC Council Workgroup Report on Behavioral Assessments, so we consider this test **PRIORITY 1.**

### Current Applications

The TMB Flicker Change Detection test is included in the Broad Neuropsychiatric Phenotyping Initiative and the 23andme cognitive testing platform. Translation of the test into standard Chinese and Spanish is currently being funded by the Broad Institute.

### Psychometric Characteristics

The primary outcome measure for Flicker Change Detection is the speed with which a participant can identify the changing portion of the image. This is calculated based on the median time (in ms) before the participant is able to identify the dot that is changing. Trials where the participant chooses an incorrect dot are not included when calculating this mean. To present a more easily interpretable outcome measure to participants, this value is also transformed into a score ranging from 0 to 25 corresponding to variations in speed.

This test has excellent internal reliability (split-half) of 0.78, as calculated from the mean reaction times of 5000 participants who completed it on TestMyBrain.

Sociodemographic effects were estimated based on the median reaction times of the 4,934 participants for whom demographic data was available. This population had a mean age of 26.59 and was 50.1% female. The distribution of scores is relatively normal (see figure 7A. Performance is variable across the life course, with reaction times decreasing (speed increasing) until approximately age 20, with slowing of reaction times throughout later adulthood (see figure 7B). Male participants show slightly faster reaction times compared to female participants (see figure 7C). Higher education is related to faster reaction times (see figure 7D).

This test shows no clear practice effects. First-time participants have a median reaction time of 6284, while repeat participants have a slower median reaction time of 6425.

### Validation

Performance on the Flicker Change Detection task is correlated with performance on other tests of change detection and visual search, as well as broader visual attention ability. This test showed moderate to high correlation with Multiple Object Tracking scores, another test of visual attention (r = 0.48, N = 10,557, 95% CI [0.47, 0.49]). It also correlated robustly with Digit Symbol Matching (spearman's rho = 0.4, N = 5777, 95% CI [0.38,0.42]) as well as a test of Visual Working Memory (r = 0.25, N = 6346, 95% CI [0.23, 0.27]). Scores on this test are also modestly correlated with tests of general cognitive ability that do not involve visual attention, such as vocabulary (r = 0.19, N = 7884, 95% CI [0.17, 0.21]).

### Appropriateness for Field Test Use

To ensure that participants understand the task, each participant completes two practice trials before the test trials begin. Thus, difficulty in comprehending the test should not pose a barrier to completion.

*Device Effects.* Reaction times on this test differ slightly between users of different digital devices (iPhone mean = 7184.77, SD = 2169.413, N = 1277; iPad mean = 8124.18, SD = 2611.95, N = 413; Macintosh laptop/desktop mean = 7015.42, SD = 2180.82, N = 1097). However, given that this test relies on very long reaction times, these differences end up being fairly small as a proportion of total variability (Cohen's d = 0.08 between iPhone and Macintosh).

*Participant Burden.* This task is relatively well-tolerated by participants. Batteries containing this test have a mean participant rating of 3.74 out of 5, close to the site-wide mean participant rating of 3.7. 86.3% of participants who begin this test complete it.

### Further Development

The requirement to press on the dot that is changing may impose limitations on smaller screens. Although we do not see this in the data, there may be counterbalancing effects of the changing dot being easier to find on small screens (smaller search space) but harder to respond to precisely - we have received complaints about responses on this test sometimes not registering on the first attempt because of this issue. In general though, this test is considered engaging and low burden.

Figure 7A. Distribution of Scores

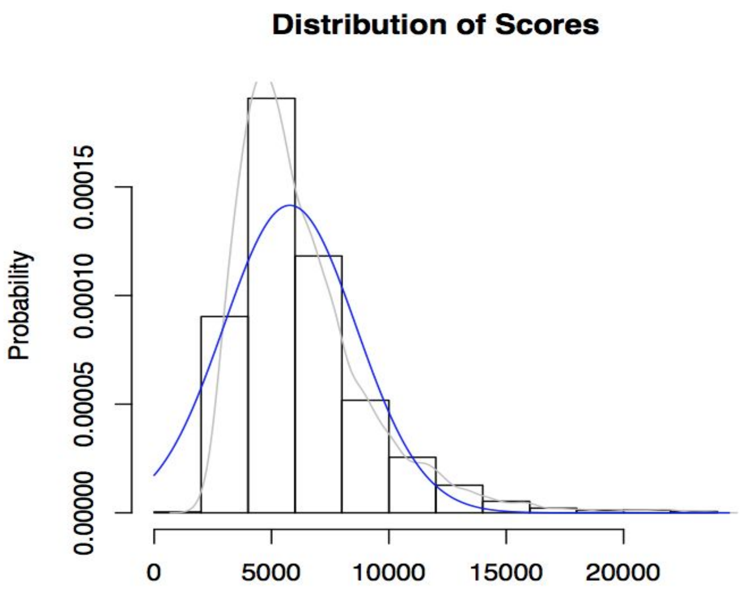


Figure 7B. Age-Related Differences in Performance (Z scores have been reversed so that better performance corresponds to higher scores)

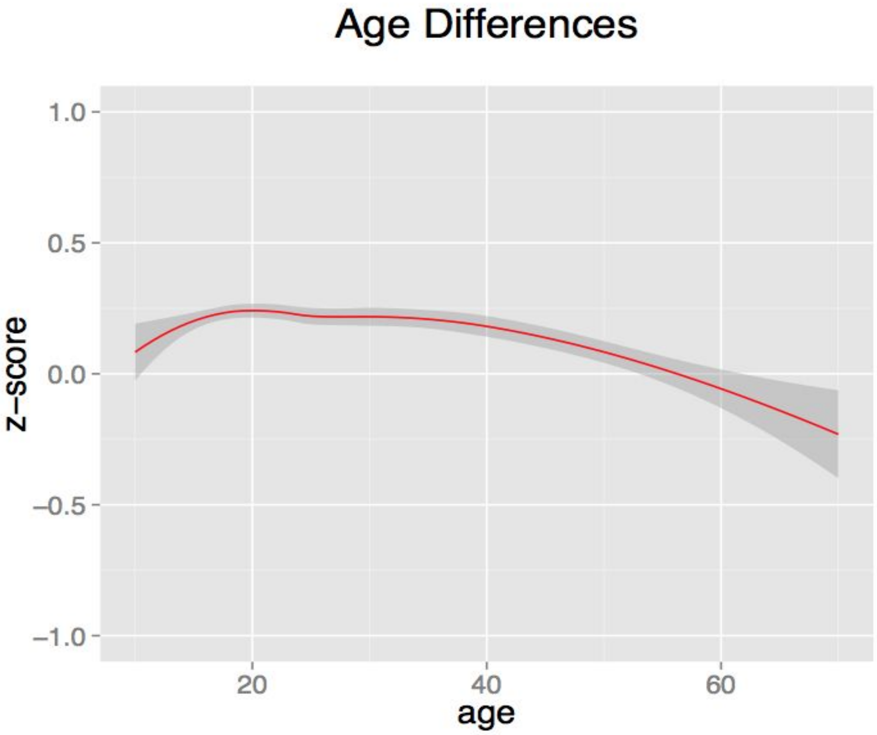


Figure 7C. Sex Differences in Performance (Z scores have been reversed so that better performance corresponds to higher scores)

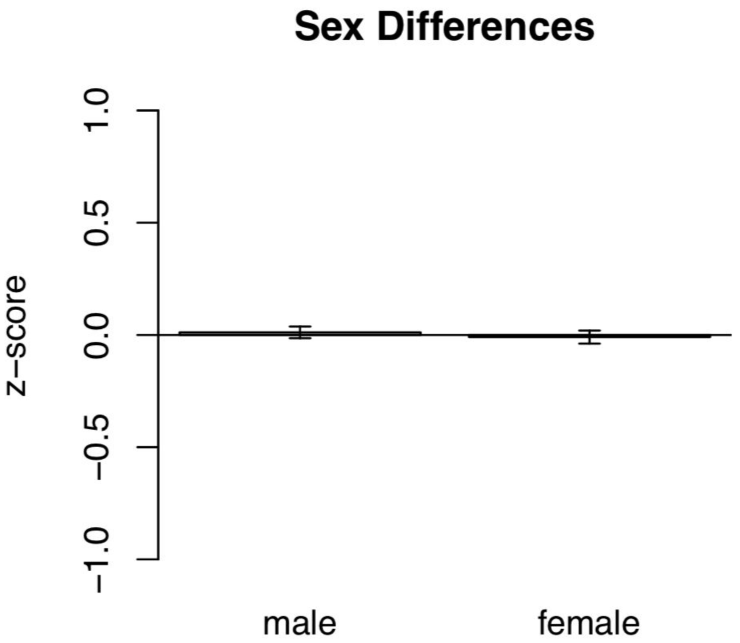
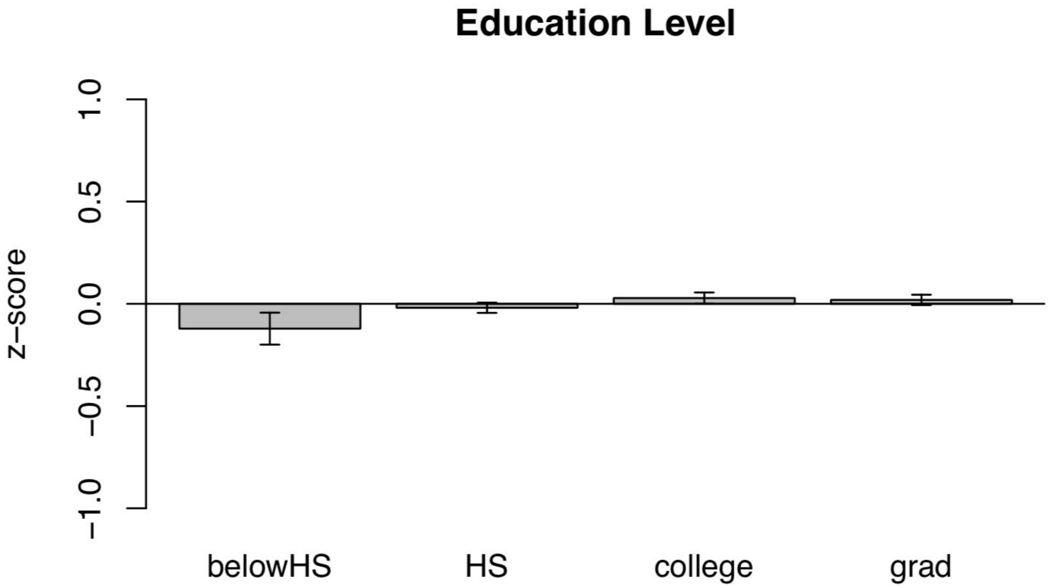
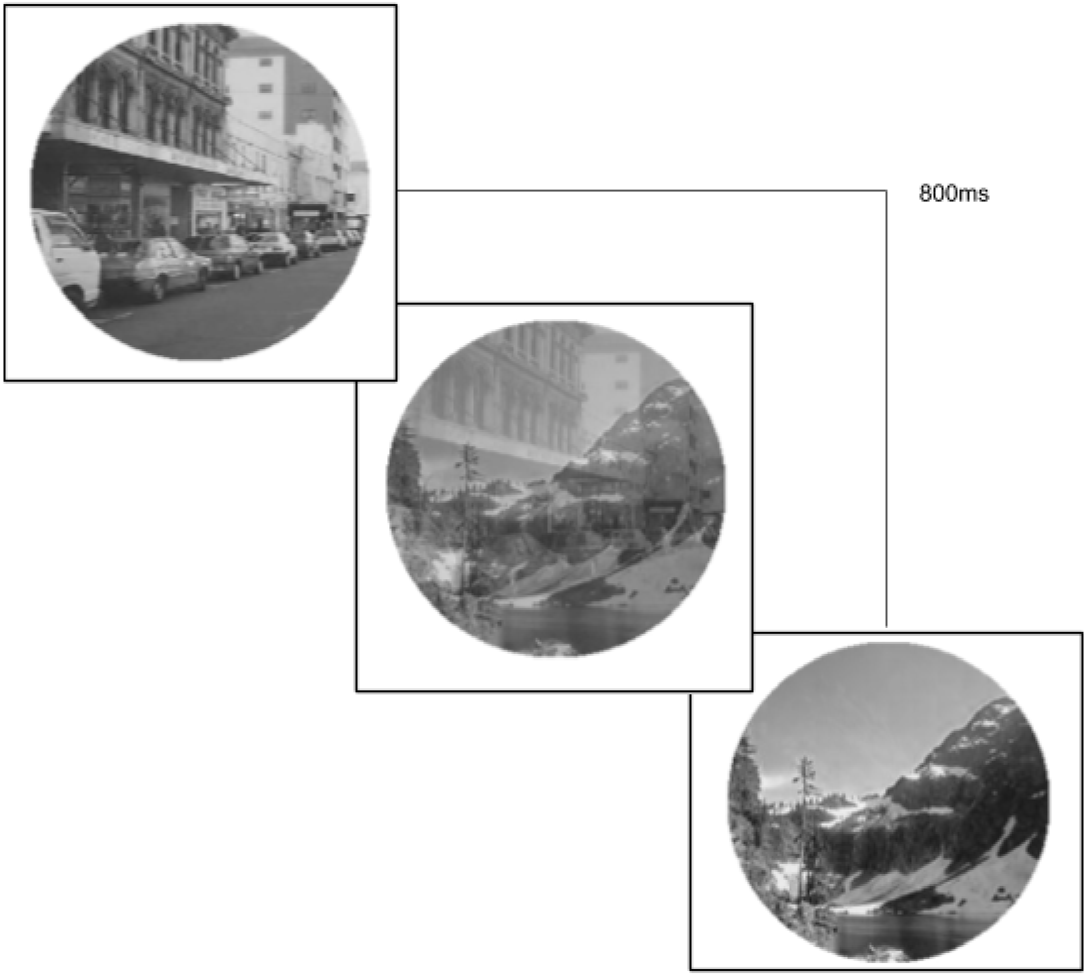


Figure 7D: Education-Related Differences in Performance (Z scores have been reversed so that better performance corresponds to higher scores)



### TMB Gradual Onset Continuous Performance Test

Constructs Measured: sustained attention, cognitive control, response inhibition
Duration: 6.9 minutes
Sample size for which normative data are available: 20,367
Demo Link: http://www.testmybrain.org/tests/gradCPT/GradCPT_Aurora1.html
Description of procedure: In this task a person attends to images that gradually transition from one image to the next. Whenever they see a street image (90% of images), they press a button. Whenever they see a mountain image (10% of images), they *do not* press a button.



This test is a standard not-X CPT (or not-X Continuous Performance Test), meaning it combines the sustained attention component of the Continuous Performance Test with a response inhibition component similar to the standard Go-No-Go test. The difference between this test and other not-X CPT variants is that this test manipulates stimulus presentation in such a way that it rapidly exhausts attentional resources, making it very sensitive to individual differences in vigilance at the same time as being very brief[1], allowing reliable measures in 3-4 minutes as opposed to 15 - 20 minutes (Rosenberg et al., 2013). In this task, participants are asked to respond to images that rapidly transition from one image to the next. Whenever they see a city scene (90% of images), they are instructed to press a button. Whenever they see a mountain scene (10% of images), they are instructed *not* to press a button. Images transition pixel-by-pixel over 800ms for a total of 4 minutes. This test was first developed by Rosenberg et al., 2013 to provide a more difficult and brief measure of sustained attention and response inhibition.

Advantages of the task are it is relatively quick and easy to administer on a range of mobile devices (particularly for this construct) and the large number of trials in a short time make it amenable to sophisticated time series modeling and analysis. Disadvantages of the task are that the nature of the construct make it high burden for most participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Both continuous performance tests and go-no-go tests are included in the RDoC Council Workgroup Report on Behavioral Assessments, so this test is considered **PRIORITY 1.**

### Current Applications

The TMB GradCPT test is currently being further developed and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project, by the 23andme personal genomics platform (adapted version), as well as in the NIMH Aurora Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

The gradCPT yields four useful measures of attention and cognitive control:

1. Omission errors - % of times the participant did not press a key when they saw a city image. More errors reflect poorer attention / concentration.
2. Commission errors - % of times the participant accidentally pressed a key when they saw a mountain image. More errors reflect poorer response inhibition.
3. Average response time - how quickly the person responded to city images. Faster response times reflect faster or more efficient information processing.
4. Response time variability - how much the person’s response time varied over the course of the task (based on standard deviation in response times). Less variable response times are associated with more consistent performance overall and are thought to reflect the ability to maintain a controlled attentional state that is colloquially referred to as being “in the zone” (Rosenberg et al., 2013).

Measures of (1) and (2) can be combined using signal detection theory to yield a *discriminability* score, reflecting the person’s ability to accurately discriminate city (key press)

[1] Most Continuous Performance Tests (CPTs) last 20 - 30 minutes.

from mountain (no key press) trials and a *bias or criterion* score, reflecting the person's response threshold or strategy used.

This test also has good reliability of 0.95 for average response time, 0.9 for response time variability, 0.77 for response sensitivity, and 0.78 for response bias (Esterman et al., 2012).

Here, we will focus on commission errors (or, accuracy on no-go trials) as the primary outcome measure or score. Based on TMB data, commission errors on the GradCPT provide both sensitive and reliable (split-half reliability of 0.7) measure, particularly given that such trials only appear 10% of the time.

Sociodemographic effects were estimated based on omission accuracy no-go accuracy (1 - commission error rate) based on a sample of 38,621 participants. The distribution of scores is relatively normal, with minor ceiling effects (see Figure 8A). Performance is variable across the lifespan, with increases in performance until about age 45 and with decreases into older age (see Figure 8B). This replicates performance patterns previously observed in participants on TMB (Fortenbaugh et al., 2015). Based on age residualized scores, there is little to no gender difference (see Figure 8C). Participants with higher levels of education are more accurate (see Figure 8D).

There are likely some practice effects on this test, but these are not evident in our database (first-time participants, no-go accuracy = 76%; repeat participants, no-go accuracy = 75%). Data from the Aurora study will allow us to quantify practice effects in the near future.

**Validation**

The gradCPT was first used by Esterman et al. (2012) as an individual differences measure of sustained attention, and performance on this task has shown to be impaired in patient populations who traditionally exhibit attention problems (DeGutis et al., 2015), correlates with self-reported attention problems in everyday life (Rosenberg et al., 2013), and fluctuates based on circadian rhythms (Riley et al., 2017). This makes it a useful task for understanding both state and trait-level differences.

Within the TMB GradCPT, commission error rate is highly correlated with variability in reaction time (coefficient of variability: standard deviation in reaction time / mean reaction time; rho = -0.44, n = 1347, 95% CIs [0.4, 0.48]). Commission errors are modestly correlated with accuracy on a Choice Reaction time test that also loads on cognitive inhibition (rho = -0.25, n = 1347, 95% CIs [-0.2, -0.3]). TMB GradCPT test performance has relatively low correlations with other distinct tests that require high effort or attention, but have a more traditional trial-by-trial structure, including the TMB Letter-Number Sequencing test (rho = -0.08, n = 2525, 95% CIs [-0.04, -0.11]) and the TMB Multiple Object Tracking test (rho = -0.11, n = 1022, 95% CIs [-0.05, -0.17]).

**Appropriateness for Field Test Use**

Considering the test is designed to be cognitively fatiguing, the TMB GradCPT test is relatively brief and reasonably well tolerated. Difficulties understanding the task (especially given the speed of the task), presents a potential challenge to completion, which has been addressed by including 3 x 1 minute practice phases before participants start the test. With this included, participants tend to know what they are supposed to do and there are minimal barriers to completion.

*Device Effects.* The TMB GradCPT test is easy to administer across a range of devices. Device characteristics are likely to impact measurements of mean reaction time, but less impactful on accuracy-based measures or measures of reaction time variability. The data show little to no effect of device type on commission error performance / accuracy (e.g. iPad mean = 70%, SD = 21%, N = 90; iPhone mean = 72%, SD = 15%, N = 106; Macintosh desktop / laptop mean = 72%, SD = 17%, N =199).

*Participant Burden.* The TMB GradCPT test is considered burdensome by participants, and is less engaging than other measures. Ratings on this test (3.6 / 5 stars) are slightly lower than the TestMyBrain.org average (3.7 / 5), with low completion rates compared with the rest of site (64% TMB GradCPT vs 81% site-wide completion among consented participants).

**Further Development**

This test can be readily modified for ecological momentary assessment designs, and we have found with measures such as d' (combining omission and commission errors) scores can be reliably obtained in 2 minutes. It would be hard to modify the test to completely reduce participant burden, however, due to the nature of the construct which depends on attentional fatigue. Otherwise, the test would be ready for field test use.

Figure 8A:  Distribution of Scores

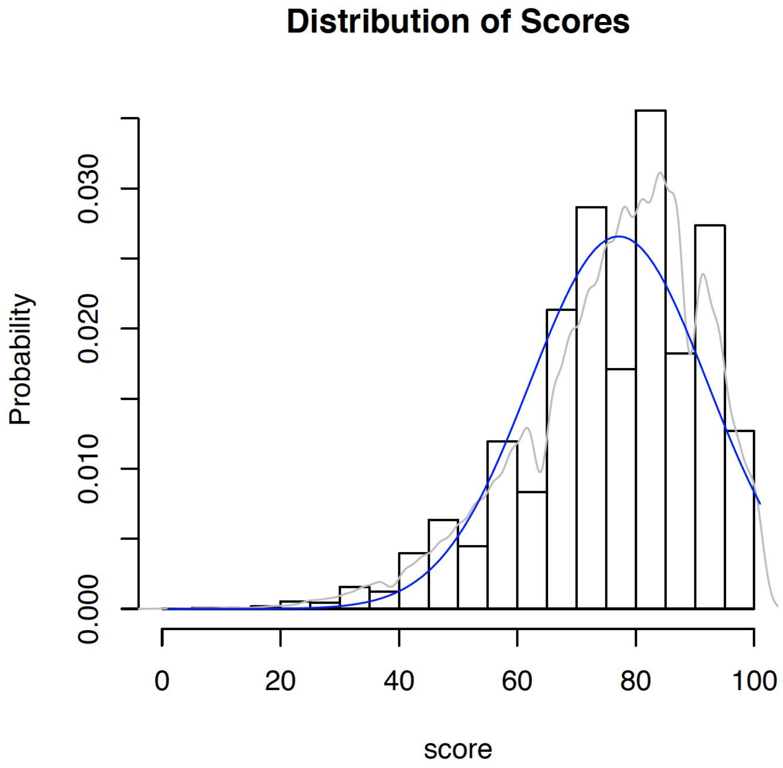


Figure 8B:  Age-related Differences in Performance

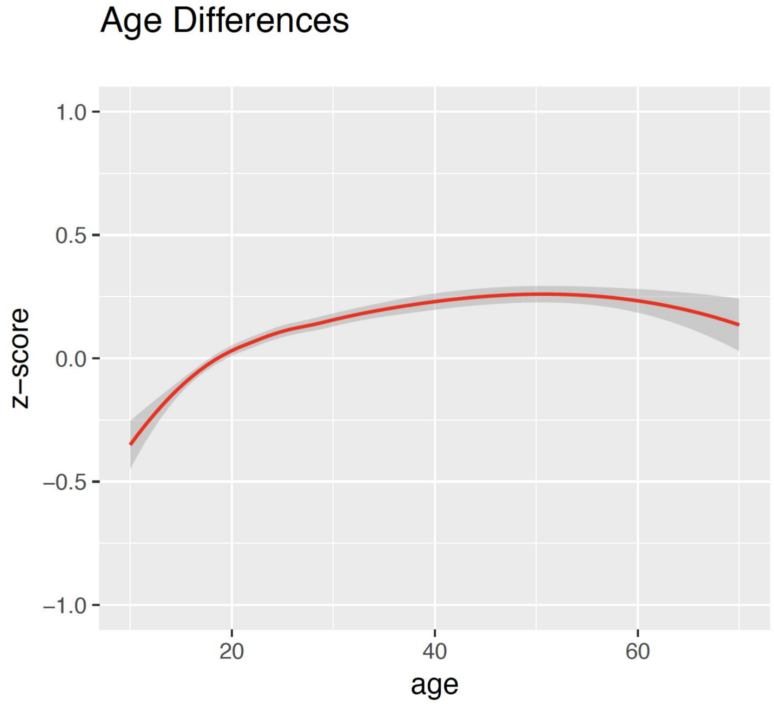


Figure 8C:  Sex Differences in Performance

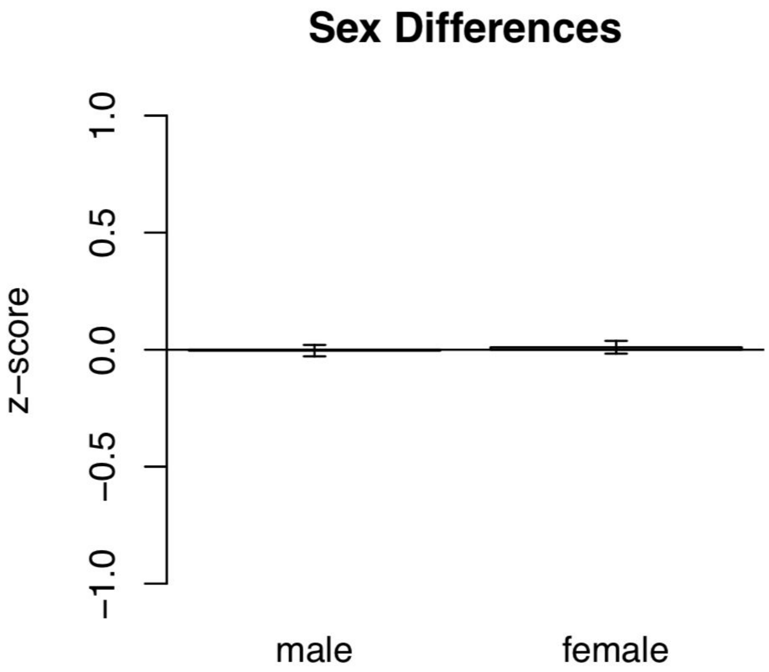
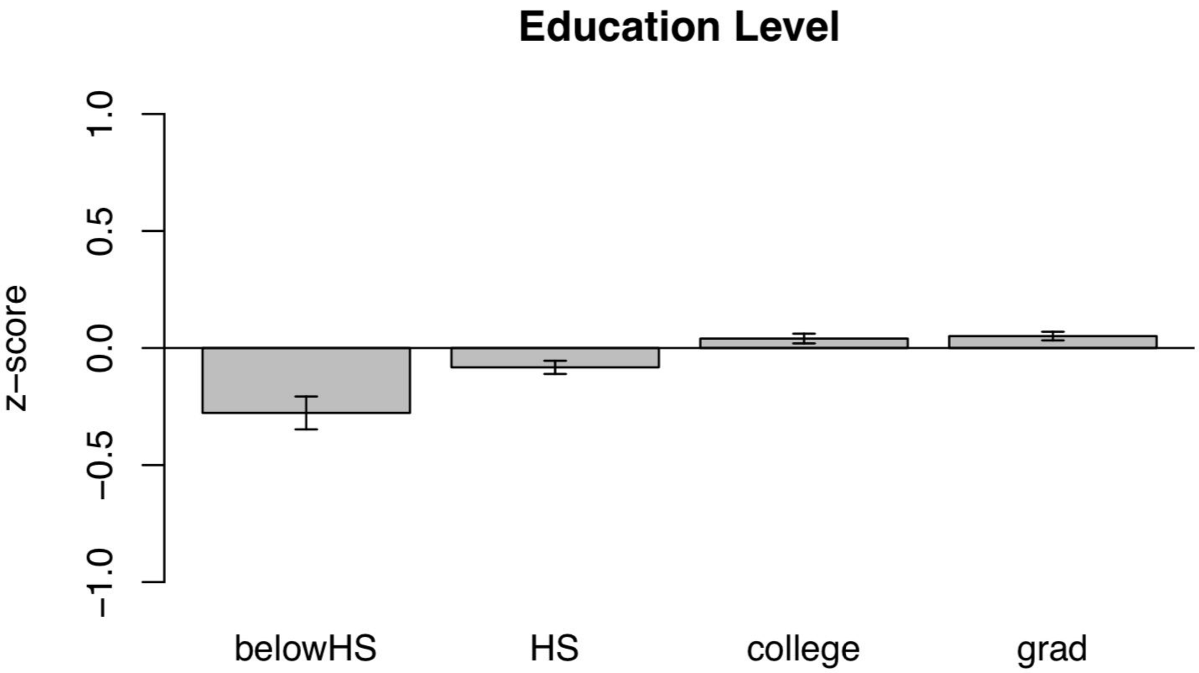
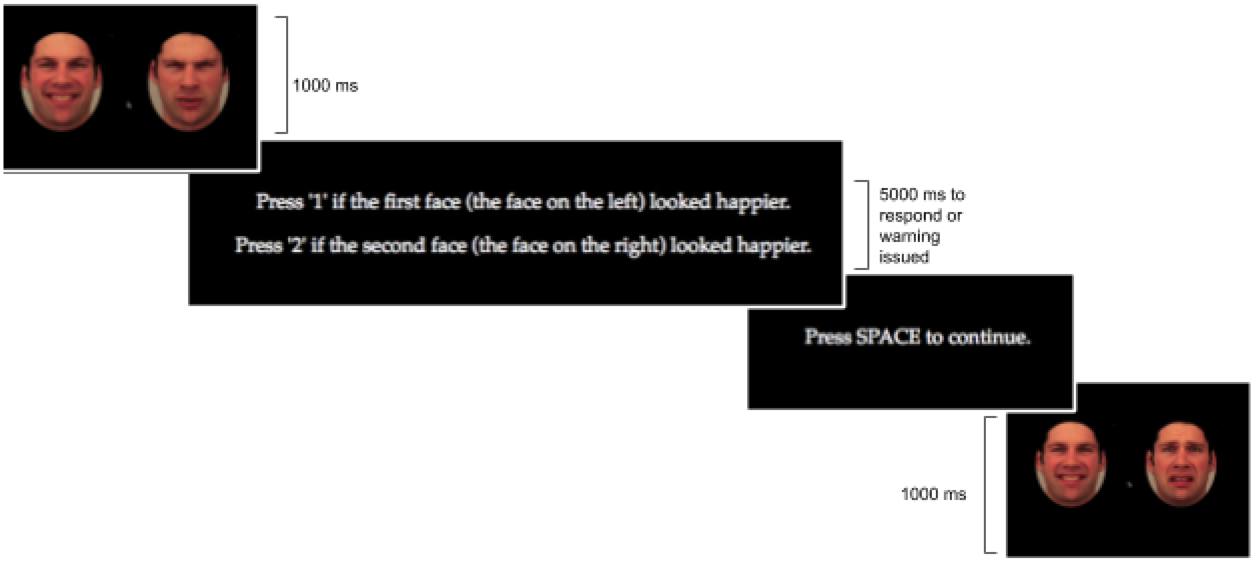


Figure 8D: Education-related Differences in Performance



### TMB Happiness Sensitivity

Constructs Measured: positive valence, social communication/reception of facial communication, understanding mental states
Duration: 2.5 minutes
Sample size for which normative data are available: 13,036
Demo Link: http://www.testmybrain.org/tests/emotion_comparison/happiness2.html
Description of procedure: Judge which of two faces is happier.



Press '1' if the first face (the face on the left) looked happier.

Press '2' if the second face (the face on the right) looked happier.

Press SPACE to continue.

This task assesses sensitivity to differences in happiness intensity, independent of response bias and differences in emotion identification or categorization (Rutter et al., 2019). Advantages of the task is it allows issues related to categorization, verbalization, response bias to be dissociated from sensitivity to specific face emotions. It is also short and easy to administer across a range of mobile device types. Disadvantages are that the task is not yet validated with respect to clinical conditions or psychopathology and is considered burdensome by participants despite its relatively short length.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Emotion sensitivity or emotion comparison tests are not included in the RDoC Council Workgroup Report on Behavioral Assessments, so we consider this test **PRIORITY 2.**

**Current Applications**

All three TMB Emotion Sensitivity tests are being used and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

**Psychometric Characteristics**

The primary outcome measure from this test is accuracy, based on proportion or number correct out of 56 trials. This score reflects the participant's ability to detect differences in happiness between faces. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures is less clear.

This test has good reliability, with internal reliability (split-half) of 0.70, as calculated from a sample of 5000 participants who completed this test on TestMyBrain.

Sociodemographic effects were estimated based on the scores of the 12,145 participants for whom demographic data was available. This population had a mean age of 27.09 and was 59.57% female. Scores are relatively normally distributed, with some ceiling effects (see Figure 9A). Performance is fairly stable across adulthood when compared to other measures: scores increase during adolescence and are consistent throughout adulthood, with a slight decline after age 60 (see Figure 9B). Female participants show slightly higher performance than male participants (see Figure 9C). Performance increases with level of education (controlled for age), though this effect is not consistent between more highly educated participant groups (see Figure 9D).

This test shows no evidence of practice effects. First-time participants had a mean score of 45.31, while repeat participants had a mean score of 44.20.

**Validation**

Performance on this test is correlated with other tests of emotion perception. It shows moderate to high correlation with performance on analogous tests of perception of anger (r = 0.46, N = 12568, 95% CI [0.45, 0.47]) and fear (r = 0.47, N = 11933, 95% CI [0.46, 0.49]). Scores are associated with current anxiety as measured by the GAD-7 (r = 0.12, N = 531, 95% CI [0.034, 0.20]), but not depression symptoms as measured by the Beck Depression Inventory (r = -0.062, N = 486, 95% CI [-0.15, 0.027]).

**Appropriateness for Field Test Use**

Before beginning the test, each participant completes 2 easy practice trials, which include immediate feedback and are repeated if the participant answers incorrectly. Therefore, difficulties in understanding the task should not present a barrier to completion.

*Device Effects.* Participants who took this test using mobile devices showed slightly lower performance than those who used laptop or desktop computers (iPhone mean = 46.86, SD = 4.07, N = 1290; iPad mean = 46.48, SD = 4.18, N = 543; Macintosh laptop/desktop mean = 47.50, SD = 3.89, N = 1239). Device type may have an impact on performance on this test; for instance, although comparable scores between iPhone and iPad suggest that screen size does not explain this difference, but may instead be due to differences in demographics or environmental context.

*Participant Burden.* This test is considered somewhat burdensome by participants. Batteries containing this test have a mean participant rating of 3.40 out of 5, compared to a site-wide mean participant rating of 3.7, however completion rates are good at 87% (compared to 81% sitewide).

### Further Development

The current version of this test relies on faces taken from predominantly Caucasian face databases, so the major limitation of this test is its use in diverse cohorts.  Versions of the test that include multiracial faces are recommended for broader applications.

Figure 9A: Distribution of Scores

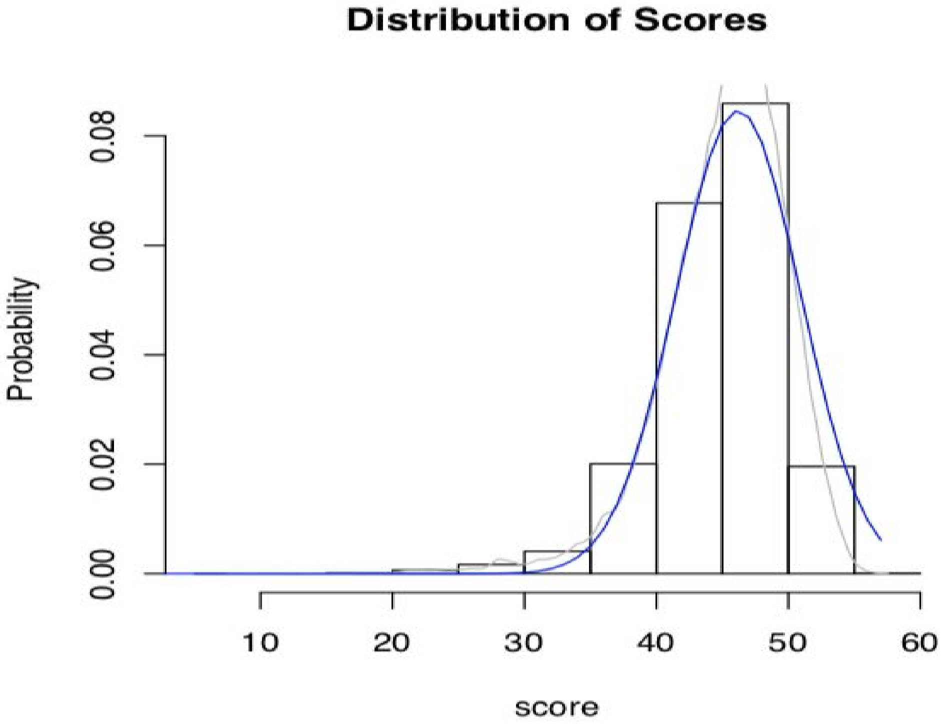


Figure 9B: Age-Related Differences in Performance

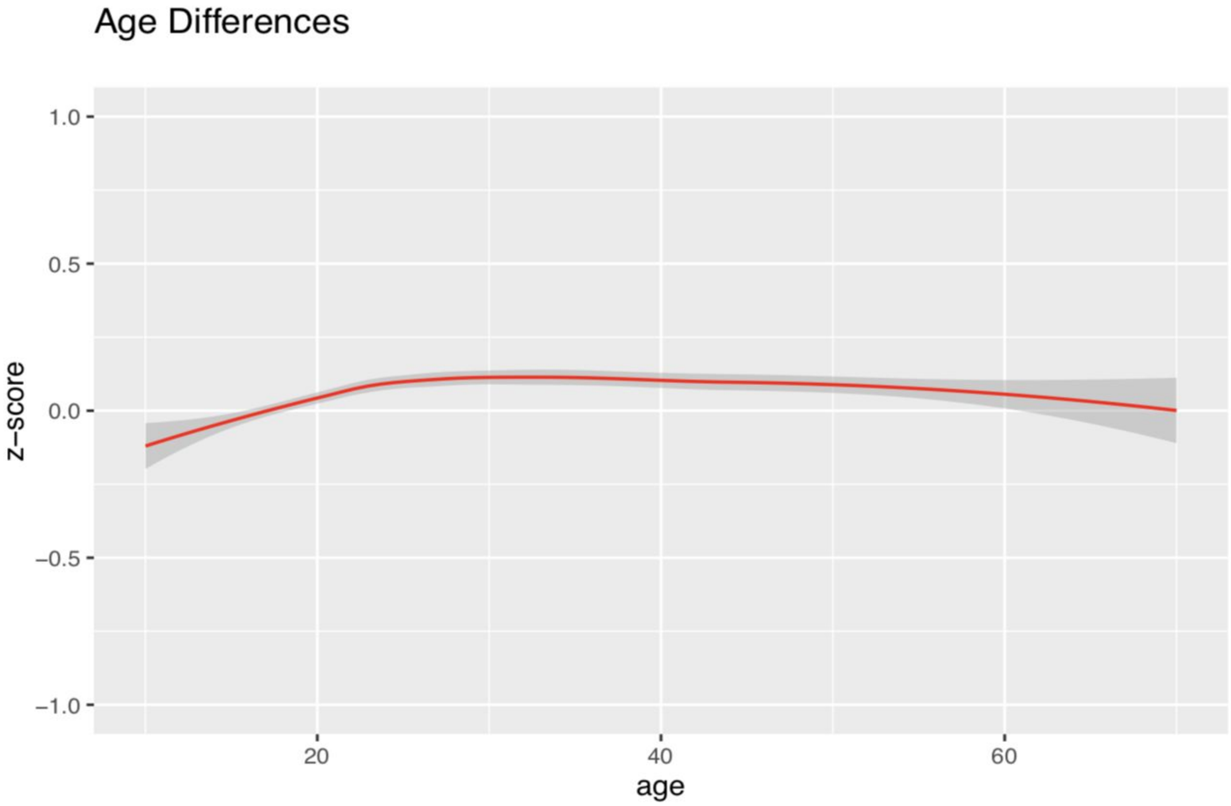


Figure 9C: Sex Differences in Performance

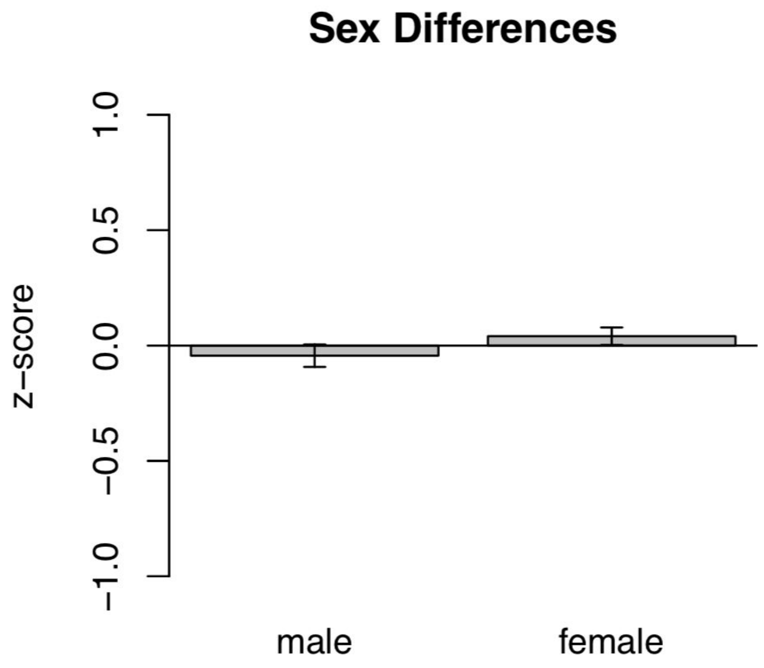
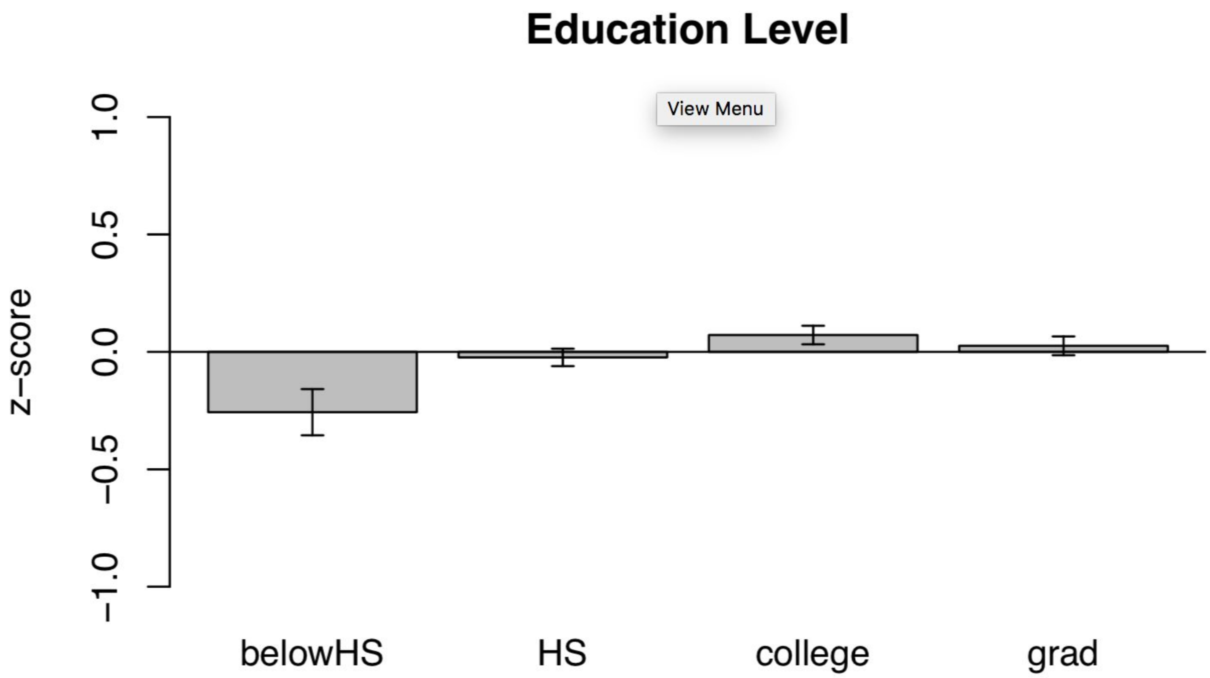
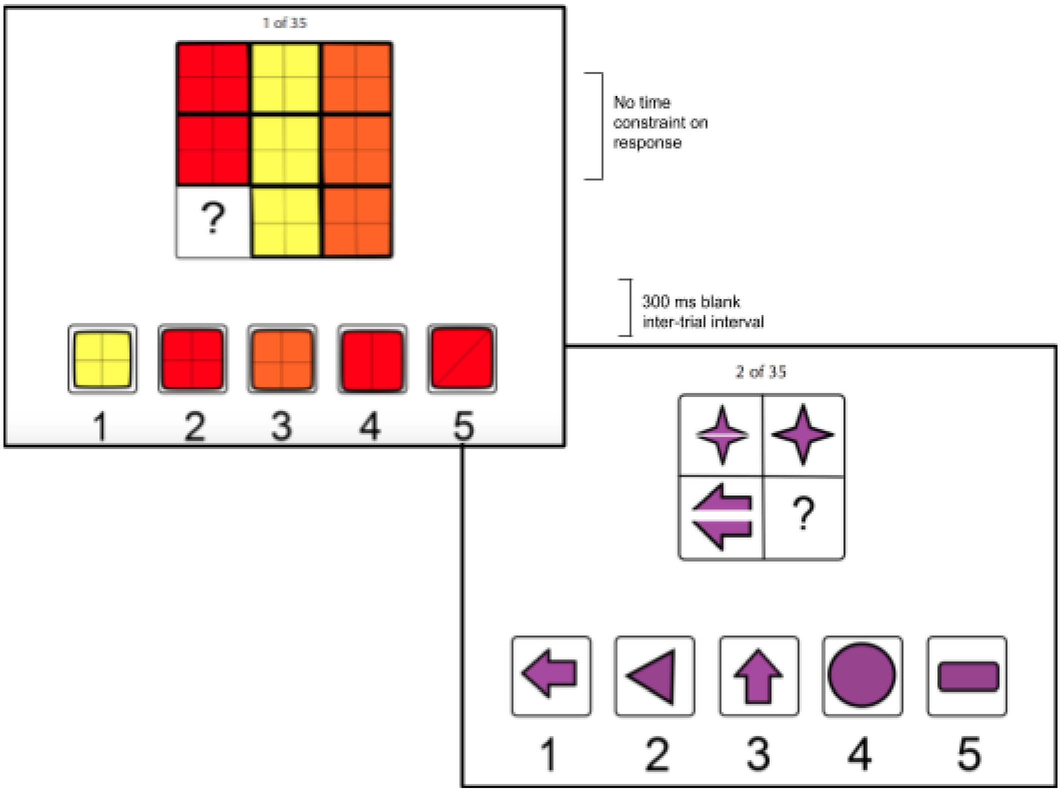


Figure 9D: Education-Related Differences in Performance



### TMB Matrix Reasoning

Construct Measures: Cognition - attention, perception, cognitive control, working memory; Also general cognitive ability, general intelligence, fluid intelligence, and nonverbal reasoning.
Duration: 8.5 minutes
Sample size for which normative data are available: 20,510
Demo Link:
http://www.testmybrain.org/tests/matrix_discontinuerule/index_neworder_discontinuerule.php
Description of procedure: Identify the image or pattern that completes an incomplete matrix of patterns, based on a logical rule.


1 of 35
?
1
2
3
4
5
No time constraint on response
300 ms blank inter-trial interval
2 of 35
?
1
2
3
4
5

This test is based on a well-validated and widely used matrix reasoning format, similar to tests that have been used in clinical neuropsychology for decades.   Advantages of the task are that it is can be administered easily on a mobile device, is considered enjoyable by participants (despite its length), can be used to measure general cognitive ability, and performance can be interpreted with respect to a large body of existing literature.  Drawbacks are specific to a field test setting and include a benefit of poor performance (task is shorter when performance is poorer due to stopping rule).

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Although nonverbal reasoning tests are not included in the RDoC Council Workgroup Report on

Behavioral Assessments, we believe that such baseline measures are important for interpreting overall performance and so have designated this task **PRIORITY 1.**

### Current Applications

The TMB Matrix Reasoning test is currently being used and evaluated as part of the Brain Health Index of the NIH Core Neuropsychological Measures for Diabetes and Obesity Project and the Broad Institute Neuropsychiatric Phenotyping Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

Here we focus on accuracy (number correct or proportion correct) as the primary outcome measure or score. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures would not be clear.

The TMB Matrix Reasoning Test has similar reliability to the original WASI II Matrix Reasoning test (Cronbach's alpha = 0.77; Spearman-Brown corrected split-half reliability, computed as in the WASI III manual, which counts all trials after its three-consecutive-incorrect stopping rule as incorrect, is 0.89).

Sociodemographic effects were estimated based on a sample of 3271 participants. The distribution of scores is relatively normal, with some ceiling effects (see Figure 10A). Performance is variable across the lifespan, with steep developmental improvements and modest age-related decline (see Figure 10B). Based on age residualized scores, there is a small gender difference that favors males (gender differences calculated on age range 18-25) (see Figure 10C) and an effect of education where participants with higher educational attainment also show superior performance (see Figure 10D).

Practice effects on this test would be considerable without the establishment of alternate forms, as once a participant figures out the rule the puzzles are easy to solve. Alternate forms would protect against such effects.

### Validation

Measures of matrix reasoning are among the best indices of fluid intelligence and also of general intelligence more broadly (Carroll, 1997). The TMB Matrix Reasoning test was modeled after the well-validated Matrix Reasoning test used in the Wechsler Abbreviated Scale of Intelligence II (Wechsler & Hsiao-pin, 2011). Matrix Reasoning tests have been used for many decades as a measure of general cognitive ability and as a "hold" test or test of "premorbid iq", since performance is relatively insensitive to variations in health in the short-term, psychological state, or many forms of brain damage (Lezak et al., 2012). For this reason, Matrix Reasoning tests provide a good control or baseline measure where measures that load on verbal ability (e.g. Vocabulary) are less desirable.

The TMB Matrix Reasoning test correlates robustly with SAT math (rho=0.41, n=1345, 95% CIs [.37, .45]); this correlation is comparable to prior reports of correlations between well-validated matrix reasoning tests and SAT math (Rohde & Thompson, 2007). As expected (Rohde & Thompson, 2007), Matrices correlates to a lesser degree, but still robustly, with SAT verbal (rho=0.22, n=1358, 95% CIs [0.17, 0.27]) and Vocabulary (rho=0.31, n=10,000, 95% CIs [0.29, 0.33]).

Controlling for participant age, the TMB Matrix Reasoning test performance correlates modestly with performance on both easy and hard versions of the TMB Vocabulary test (30 item easy: rho = 0.29, n = 1686, 95% CIs [0.25, 0.33]; 20 item hard: rho = 0.35, n = 1511, 95% CIs [0.31, 0.39]) and well as the TMB Digit Symbol Matching test (processing speed) (rho = 0.38, n = 1210, 95% CIs [0.33, 0.43]).

**Appropriateness for Field Test Use**

Overall, the TMB Matrix Reasoning test is an interesting and engaging test for participants with minimal technical barriers. Practice items and increasing difficulty from the beginning to the end of the test ensure that participants know what they are supposed to do and there are minimal barriers to completion.

*Device Effects.* The TMB Matrix Reasoning test is relatively easy to administer across a range of devices. With some items, there may be a concern that stimuli are too complicated to perceive accurately on smaller screens, but the data do not clearly reflect this (e.g. iPad mean = 27.4, SD = 4.2, N = 1561; iPhone mean = 27.2, SD = 4.2, N = 1854). There is an effect of portable vs. nonportable device type that probably indicates differences in administration environment (Macintosh desktop / laptop mean = 28.8, SD = 3.8, N = 3788), although further analyses would be needed to better understand these differences.

*Participant Burden.* The TMB Matrix Reasoning test is considered enjoyable by participants despite its length. Ratings on this test (3.83 / 5 stars) compare favorably with average ratings on TestMyBrain.org (3.67 / 5), despite its relatively long length (avg = 8.7 minutes), with completion rates that are higher than the rest of the site (90% TMB Matrix Reasoning vs 81% site-wide completion among consented participants).

**Further Development**

The most obvious next step for development of the TMB Matrix Reasoning test for field test use is to create an Item Response Theory (IRT) adaptive version of the test. The independence of individual item performance combined with varying levels of difficulty mean that IRT is both appropriate for this test, and might reduce the length of the test considerably.

Figure 10A. Distribution of Scores

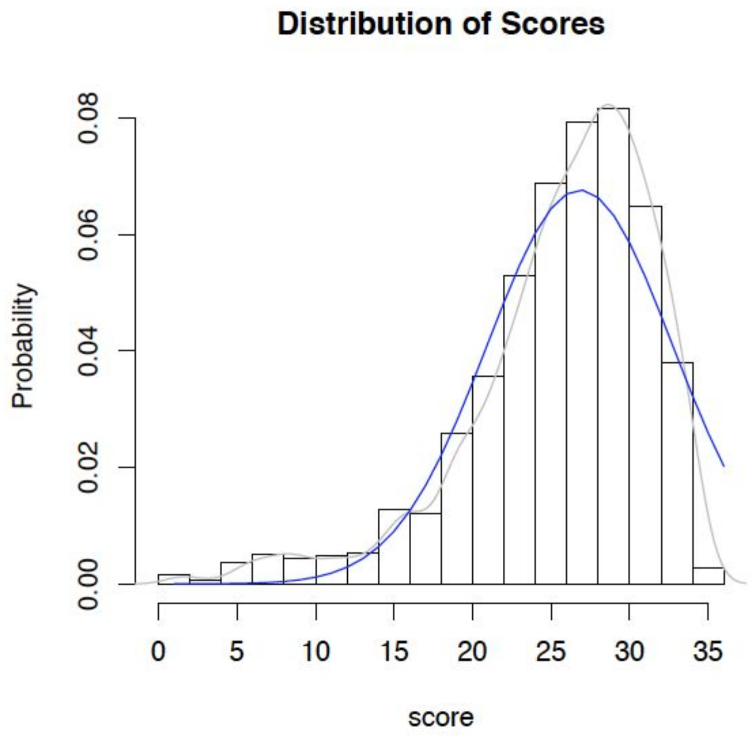


Figure 10B. Age-Related Differences in Performance

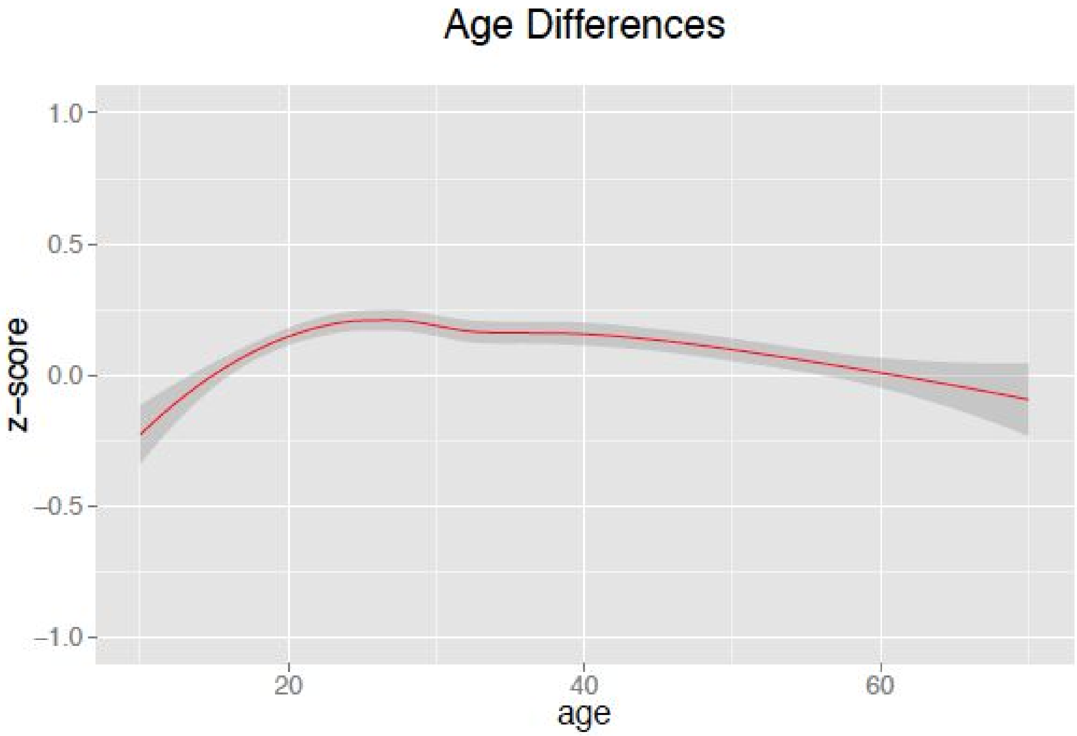


Figure 10C. Sex Differences in Performance

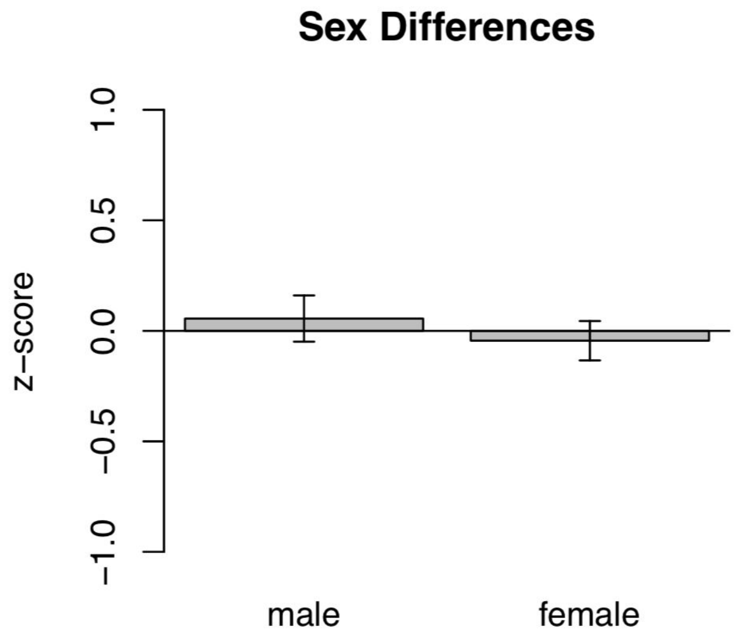
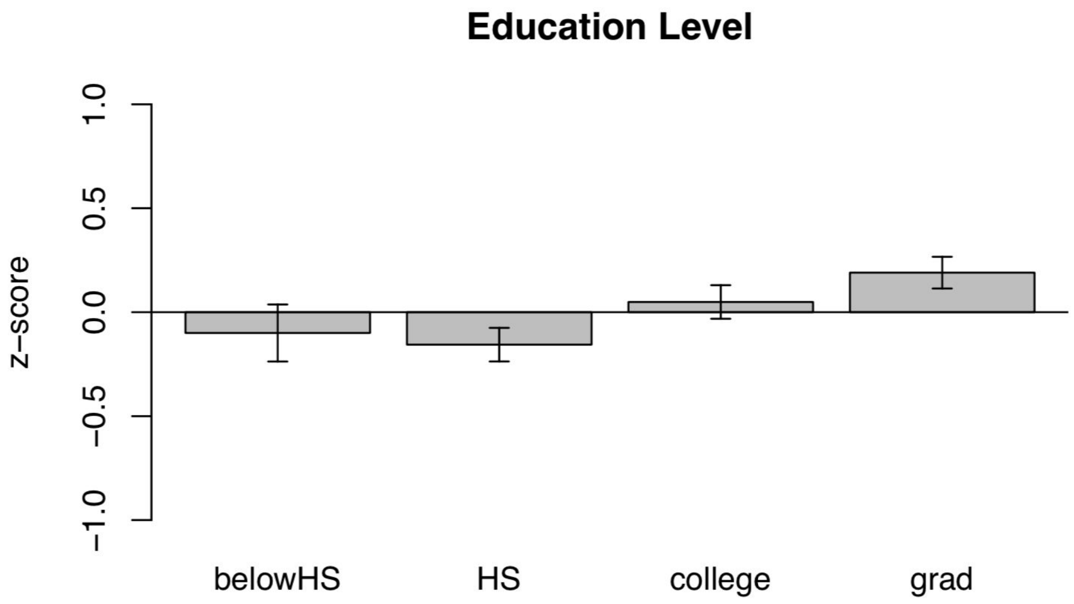
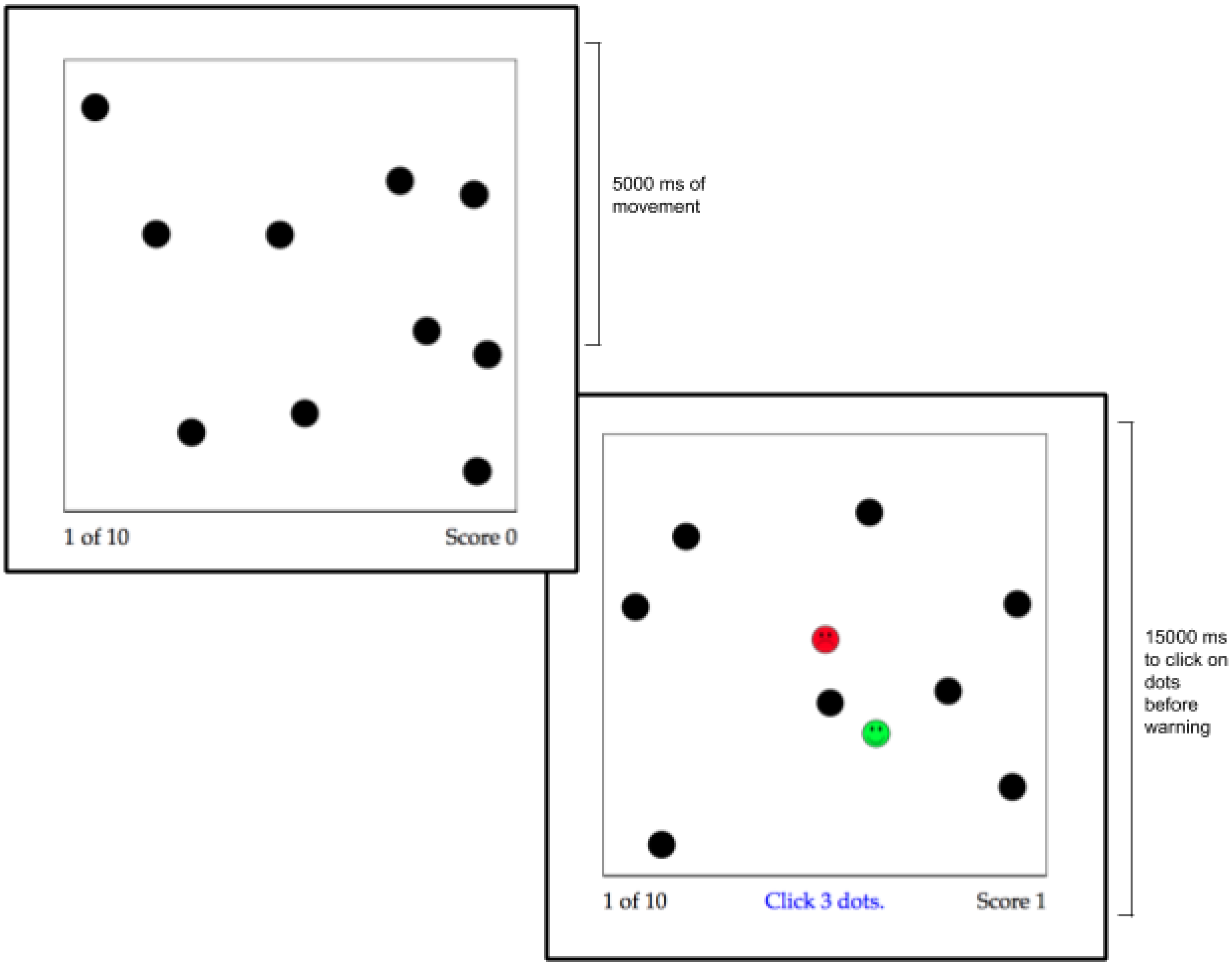


Figure 10D. Education-related Differences in Performance



### TMB Multiple Object Tracking

Construct Measures: Cognition: Attention, Perception, Cognitive Control, Working Memory
Duration: 10 minutes (standard length); shorter versions available
Sample size for which normative data are available: 43,805
Demo Link:
http://www.testmybrain.org/tests/multiple_object_tracking_v1_3/MOTTest_SATm2.html
Description of procedure: Remember and track a set of target circles as they move around the screen, among a larger set of identical distractor circles.



This is a standard multiple-object tracking task that measures visuospatial attention and short-term memory (Wilmer et al., 2016). The task varies both the number and speed of dots that need to be tracked. Advantages of the task that it is available in short format (5 min) and is considered very engaging by participants. It also dissociates almost entirely from sustained attention, giving it interesting psychometric characteristics for a demanding cognitive test. It can be administered quickly and easily on a range of mobile devices. Disadvantages are that the task relies on complex stimuli that may be affected by future changes in device displays. There are also may be potential limitations in stimulus delivery on a smaller screen.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Visual object tracking is not included in the RDoC Council Workgroup Report on Behavioral Assessments, however, so this test is considered **PRIORITY 2.**

### Current Applications

The TMB Multiple Object Tracking test is currently being used and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

The primary outcome measure for this test is the number of dots that a participant was able to track and identify successfully (a score ranging from 0 to 120). There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures would not be clear.

This test shows excellent reliability; internal reliability (split-half) was 0.92, calculated from the scores of 5000 participants who completed this test on TestMyBrain.

Sociodemographic effects were estimated based on the scores of 6882 participants for whom demographic data was available. This participant group had a mean eage of 28.92 and was 52.48% female. The distribution of scores is relatively normal, but shows some ceiling effects (see Figure 11A). Performance is variable across the lifespan, increasing throughout adolescence and young adulthood and peaking at approximately age 25 before declining throughout adulthood (see Figure 11B). Male participants show higher scores, on average, than female participants (see Figure 11C). Performance increases with education, though this effect is not apparent when comparing the more highly educated participant groups (see Figure 11D).

Practice effects on this test are minimal. The mean score for first-time participants was 80% correct, while the mean score for repeat participants was 81% correct (cohen's d = 0.1).

### Validation

Performance on the Multiple Object Tracking test correlates with other tests of attention, cognitive control, and working memory (all correlations are controlled for age where age data was available). This test showed moderate to high correlation with Flicker Change Detection, another test of visual attention (r = 0.48, N = 10,557, 95% CI [0.47, 0.49]). It showed much lower correlation with the Gradual Onset Continuous Performance Task, a test of sustained attention (r = 0.063, N = 1066, 95% CI [0.0032, 0.12]), as well as with self-reports of impaired attention (r = -0.003, N = 813, 95% CI = [-0.071, 0.066]). It also shows moderate to high correlation with Matrix Reasoning, a test of visual pattern recognition (r = 0.51, N = 84, 95% CI [0.33, 0.65]; this correlation could not be controlled for age). However, it does not correlate significantly with other tests of memory, such as vocabulary (r = .0093, N = 95, 95% CI [-0.19, 0.21], not age-adjusted) or forward digit span (r = 0.0053, N = 34, 95% CI [-0.33, 0.34]). Notably, scores on this test are well correlated with math SAT scores (r = 0.27, N = 3,304, 95% CI [0.24, 0.30]) and less so with verbal SAT scores (r = 0.1, N = 3,329, 95% CI [0.07, 0.13]).

### Appropriateness for Field Test Use

In order to ensure that participants understand the task presented to them, the test includes two practice trials that give direct feedback to participants before test trials begin. Thus, difficulty in understanding the test should not present a barrier to completion.

*Device Effects.* Users of all device types perform at similar levels on this test (iPhone mean = 79%, SD = 10%, N = 526; iPad mean = 79%, SD = 11%, N = 404; Macintosh laptop/desktop mean = 80%, SD = 10%, N = 1426), with slightly higher performance on laptop or desktop computers than users of mobile devices. Given the complex nature of the visual stimulus, cautioned should be used, however, before administering these tests on mobile devices with small screens.

*Participant Burden.* This test was rated as highly engaging by participants (3.9 / 5 vs. 3.7 / 5 for other tests), although completion rates are only modestly higher than average (85% vs. 81% for other tests), likely due to the test's long length relative to other tests.

### Further Development

Given its high reliability, this test can likely be shortened considerably (e.g. from 9 min to 2.5 min) and still maintain acceptable reliability (estimated internal reliability = 0.72). Also, the fact that trials vary in difficult means that test performance can also be estimated using item response theory (IRT) models, allowing discrepant patterns of performance that indicate poor validity to be identified.

Figure 11A. Distribution of Scores

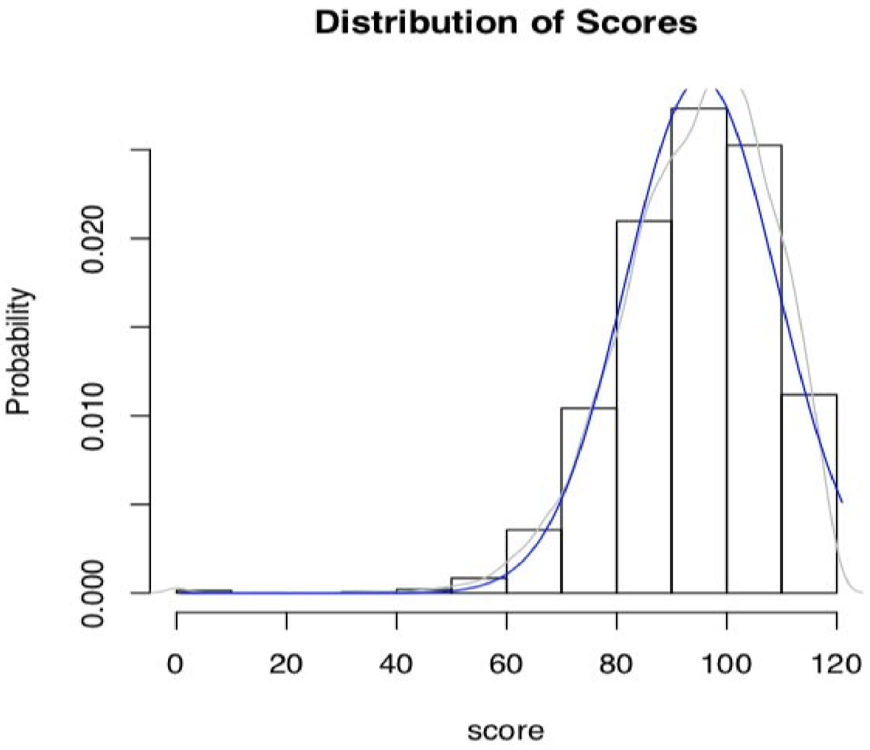


Figure 11B. Age-Related Differences in Performance

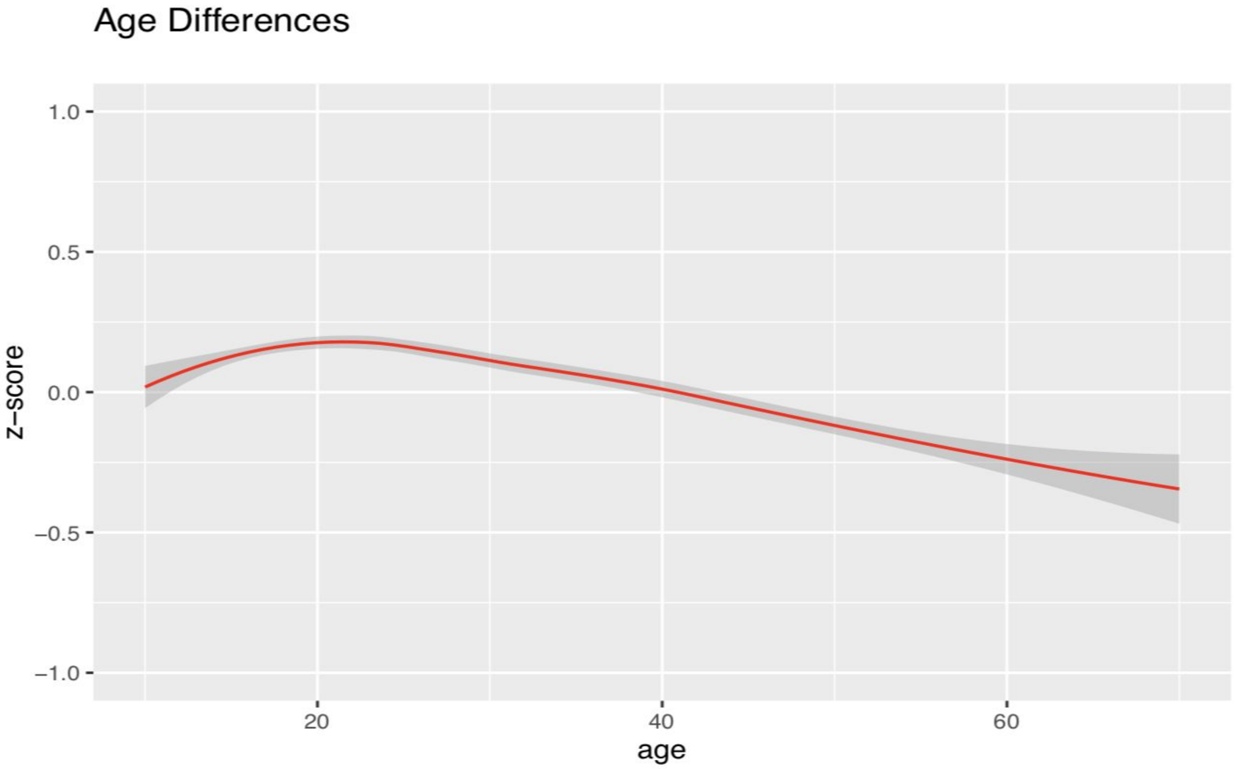


Figure 11C. Sex Differences in Performance

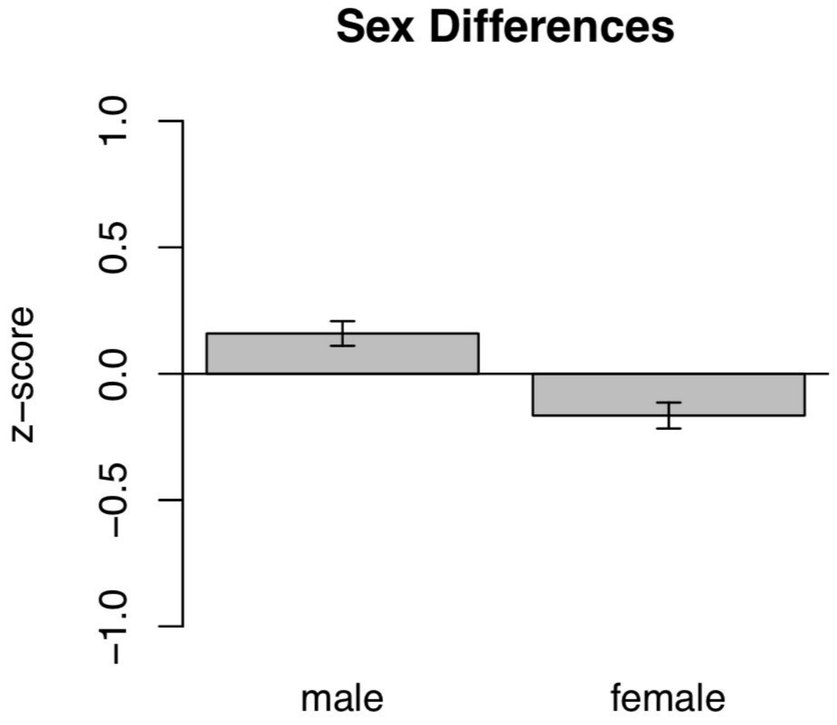
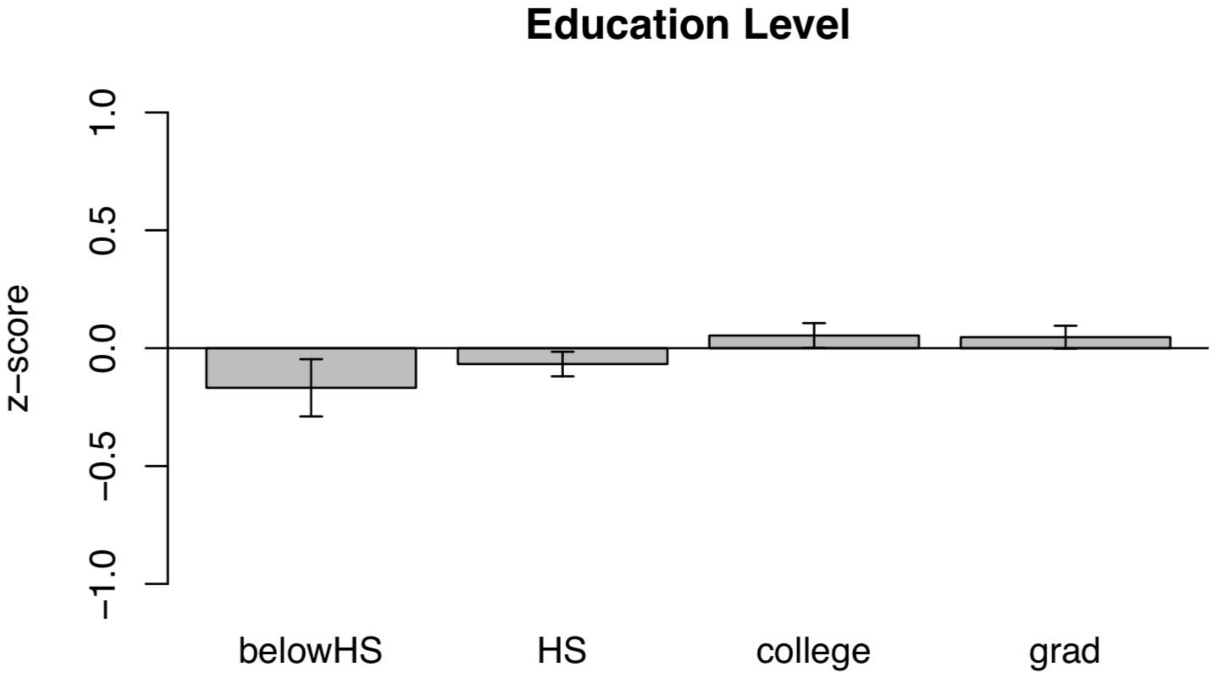
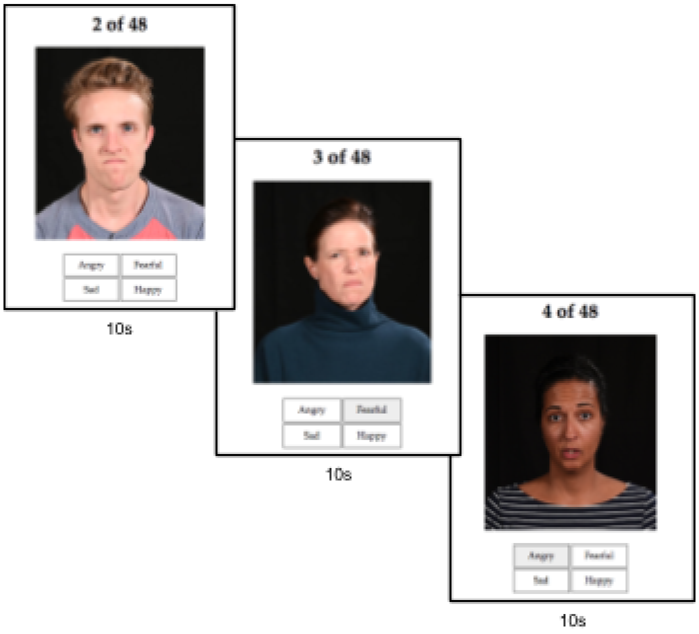


Figure 11D. Education-Related Differences in Performance



### TMB Multiracial Face Emotion Identification Test

Construct Measures: Social Communication / Reception of Facial Communication
Duration: 2.6 minutes
Sample size for which normative data are available: 14,332
Demo Link: http://www.testmybrain.org/tests/aobh_emotion/id_emotions4.php
Description of Procedures: Identify which of four emotions (anger, happiness, fear, and sadness) best describes the emotion in a face.



This test is a standard format basic emotion identification test that was designed using item analysis off a larger item bank to increase difficulty. In this task, the participant is asked to identify which of four basic emotions best describes a series of 47 faces (angry, happy, sad, or fearful). Faces represent a broad range of adult ages and race/ethnicities, with approximately equal proportions of men and women (Dodell-Feder, Ressler & Germine, in press). This is a novel test that is based on a standard format basic emotion identification test, such as the Penn ER-40, but without a neutral condition (due to psychometric reasons; see Task Development below). Advantages of the task are that it is short, can be administered quickly and easily on a mobile device, and includes faces across a range of ages and ethnicities as part of the Act Out for Brain Health project at the Company One theater.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. This test was adapted from a similar emotion identification task (the Penn ER-40) which is included in the

RDoC Council Workgroup Report on Behavioral Assessments, so we consider this test **PRIORITY 1.**

### Task Development

To develop this task, we recruited actors from across a range of ages and race/ethnicities, from the Boston Company One theater, as part of the Act Out for Brain Health project. Boston Company One theater has a mission of engaging the city's diverse communities, with an emphasis on diverse actors. Images were taken from video clips of actors portraying different emotions. An initial set of 146 images were selected portraying anger, fear, happiness, sadness, and neutral facial expressions to create an item bank. Images were drawn from this item bank and data was collected from a development sample of N= 8309 participants who each saw a subset of 37 - 53 images. Ultimately the neutral condition was dropped as these faces were judged with significantly (ps < 0.01) poorer reliability than anger, fear, sadness, and happiness (average correlation with rest of items for each emotion category: anger: r=0.3; fear: r = 0.26; sadness: r = 0.2; happiness: r = 0.25; neutral: r = 0.06). The reliability of judgements of other emotions did not significantly differ from each other (ps > 0.1). Overall, performance on valenced items (anger, sadness, fear, happiness) captured substantially more variance in total scores than neutral items ($R^2$ 4-9% for anger, fear, happy, neutral and $R^2$ 0.4% for neutral). The final test includes 48 images that were selected to capture (1) images with consistent judgments of a single emotion, (2) varying levels of difficulty for each emotion, and (3) items with high correlations with overall emotion recognition accuracy to maximize reliability, while preserving the diversity of actors and faces.

### Current Applications

The TMB Multiracial Emotion Identification Test test is currently being further developed and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project as well as in the NIMH Aurora Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

Here we focus on accuracy (number correct or proportion correct) as the primary outcome measure or score. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), as well as performance on individual emotion categories. These can be examined more specifically if desired.

The TMB Multiracial Face Emotion Identification test has a Cronbach's alpha of 0.75, which compares favorably with other tests of the same form.

Sociodemographic effects were estimated based on a sample of 13,174 participants. The distribution of scores is relatively normal, with minor ceiling effects (see Figure 12A). Performance is variable across the lifespan, with increases in performance until about age 30 and minimal decreases into older age, consistent with other tests of similar constructs (Hartshorne & Germine, 2015) (see Figure 12B). Based on age residualized scores, there is a sex difference in performance, with female participants performing better than male participants (see Figure 12C). Participants with higher levels of education had better performance, but the difference was not consistent across education levels (see Figure 12D).

Practice effects are likely over a time interval where faces might be remembered from previous testing. Therefore, alternate forms would be recommended in situations where retest is considered.

### Validation

Emotion identification tasks are widely used in the neuropsychiatric literature as a way of estimating social cognitive ability or identifying social cognitive or social perceptual impairments. Such impairments are consistently identified, particularly in individuals with severe neuropsychiatric disorders such as schizophrenia, bipolar disorder, and autism spectrum disorders, as well as in individuals at high risk of such disorders.

For this particular test, overall performance is modestly correlated with performance on the Reading the Mind in the Eyes test (r = 0.35, N = 160, 95% CIs [0.21, 0.48]) as well as the TMB Vocabulary test (r = 0.29, N = 1141, 95% CIs [0.24, 0.34]). Performance is also associated with levels of social anhedonia in the population, a risk factor for psychosis (r = 0.11, N = 6717, 95% CIs [0.09, 0.13]) as well as individual differences in prodromal symptoms (r = 0.12, N = 8213, 95% CIs [0.1, 0.14]). These effect sizes are comparable to those reported in the literature between psychosis risk and traditional, well-validated emotion identification tests (Germine & Hooker, 2011; Germine et al., 2011).

### Appropriateness for Field Test Use

This test is very brief and is considered engaging by participants (based on attrition and ratings combined), with minimal technical or user interface barriers to completion. We would consider it ready for field test use.

*Device Effects.* The TMB Multiracial Face Emotion Identification test is easy to administer across a range of devices. The data show little to no effect of device type on accuracy (e.g. iPad mean = 40.5, SD = 4.5, N = 807; iPhone mean = 40.6, SD = 4.2, N = 1069; Macintosh desktop / laptop mean = 40.8, SD = 5.1, N =1990).

*Participant Burden.* The TMB Multiracial Face Emotion Identification test is given an average rating by participants (3.7 / 5 stars, comparable to ratings sitewide), but is completed at a very high rate relative to other tests (97% vs. 81%). Overall, it is low burden relative to other measures.

### Further Development

Creation of alternate forms with comparable reliability would be  a useful next step for development of this test. Otherwise, it is currently appropriate for a field test battery.

Figure 12A. Distribution of Scores

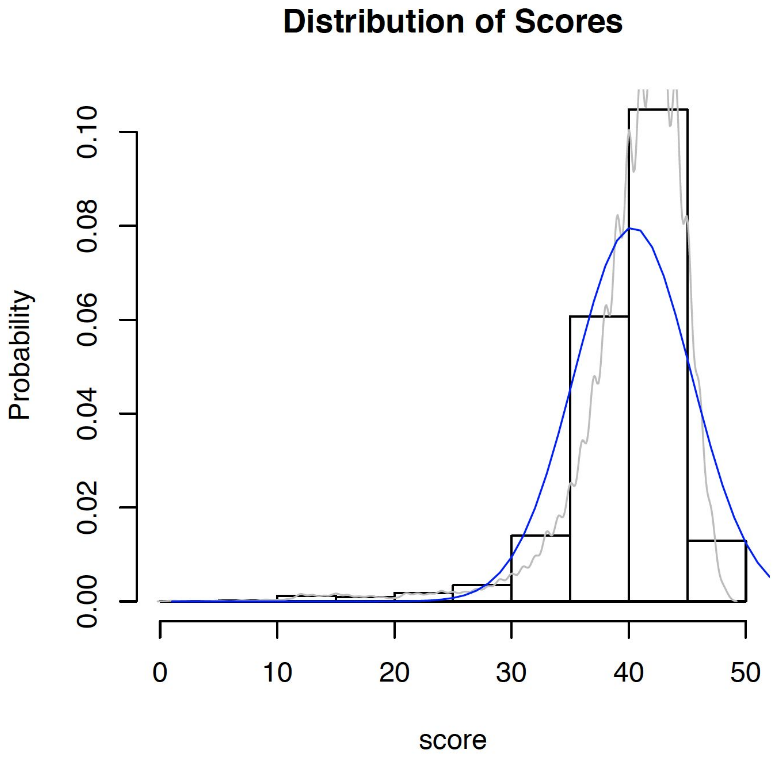


Figure 12B. Age-related Differences in Performance

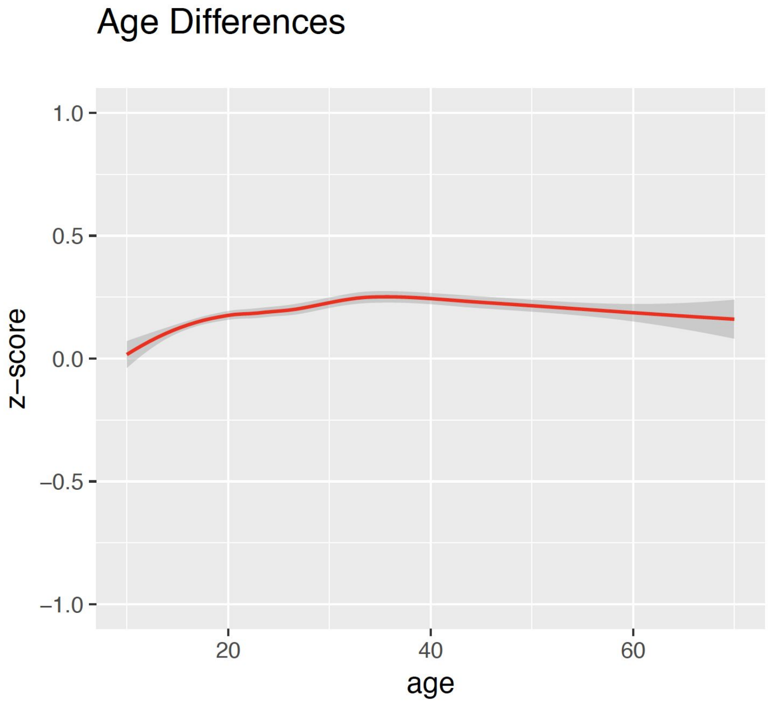


Figure 12C. Sex Differences in Performance

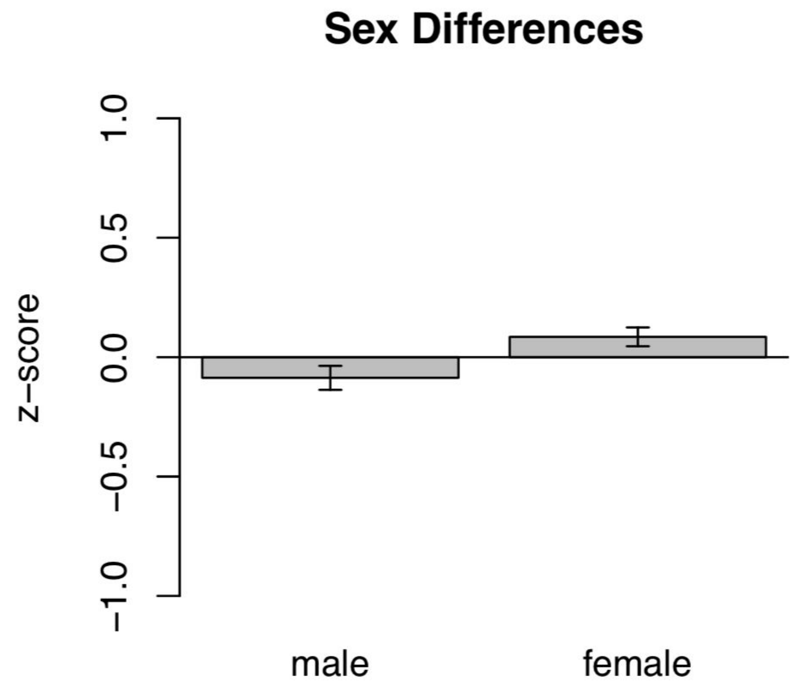
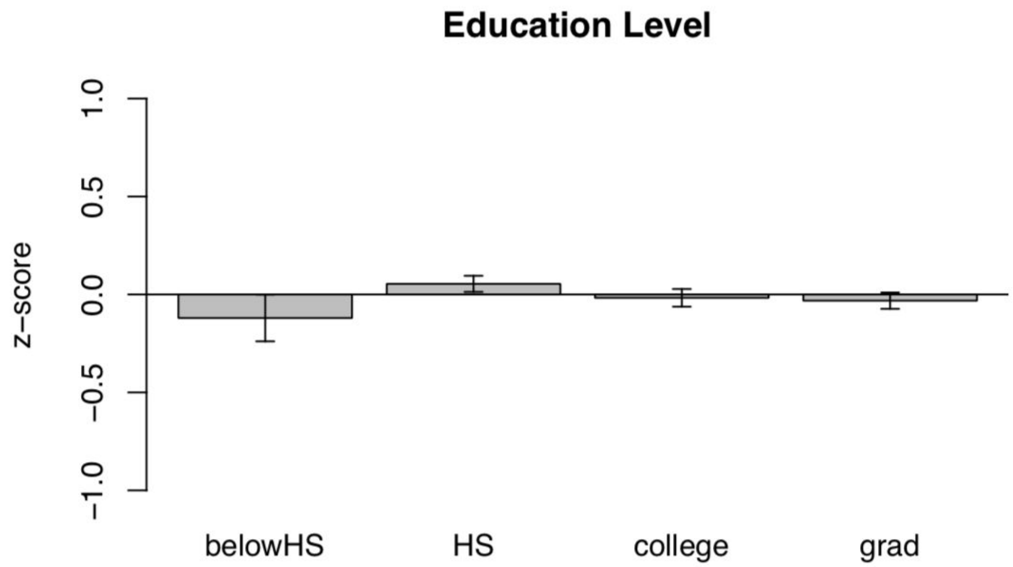
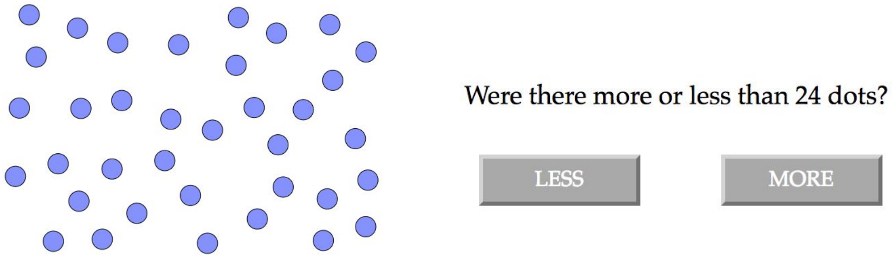


Figure 12D. Education-related Differences in Performance



**TMB Probabilistic Reward Test**

Construct Measures: Reward Learning
Duration: 9.5 minutes
Sample size for which normative data are available: 130
Demo Link: http://www.testmybrain.org/tests/prt/prt_bait_onecolor.html
Description of procedure: Decide whether there are more or less than a target number of dots on the screen (Figure 1). The perceptual discrimination is difficult, and one of the two responses (LESS or MORE) is rewarded more than the other—this leads to a response bias towards the more frequently rewarded option.



Were there more or less than 24 dots?

This test is based on a well-validated and widely used measure of probabilistic reward learning (Pizzagalli et al., 2005). Advantages of the task are that it can be administered on a mobile device and performance can be interpreted with respect to a large body of literature. Initial versions, however, did not produce a reliable response bias across a range of devices, limiting usefulness in a field test battery. After some trial and error (described below), we were able to identify a format that reliably produces response bias on mobile devices. A disadvantage of the task is that it is fairly long and viewed as burdensome by some participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. The probabilistic reward task is included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 1.**

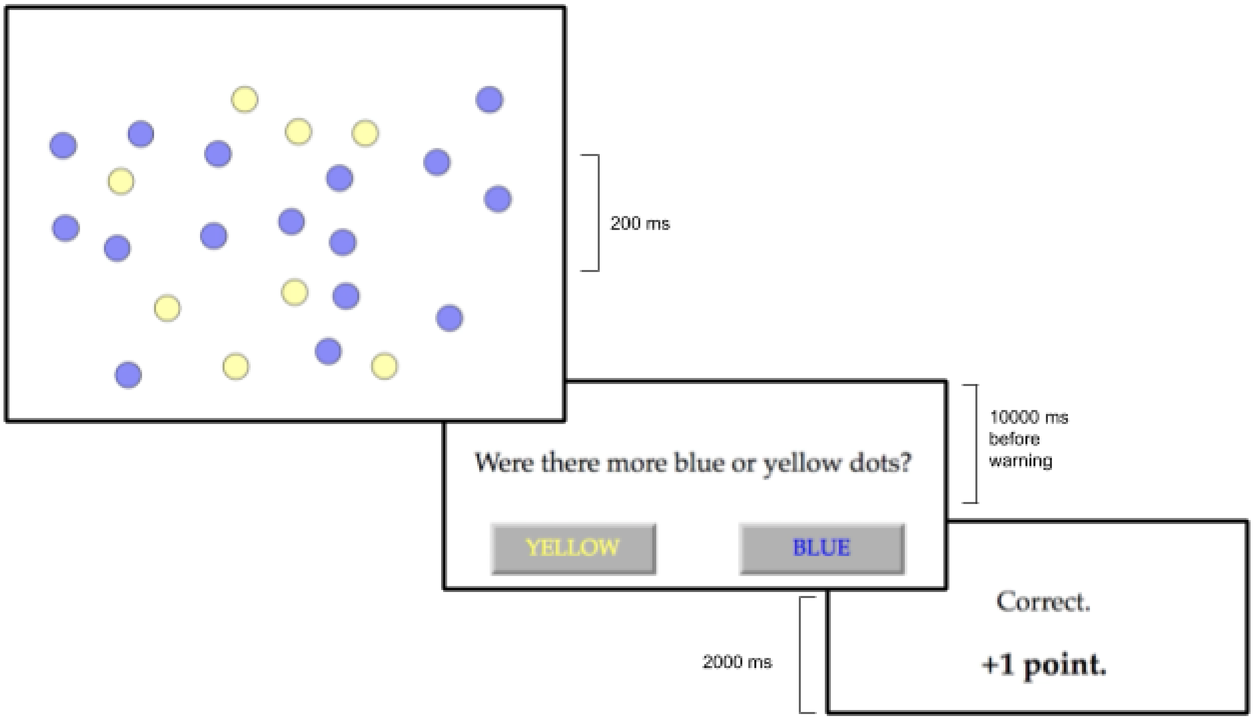
### Current Applications

A version of this test is currently being used in the NIMH Aurora Study, although we plan to replace it with the version described here.

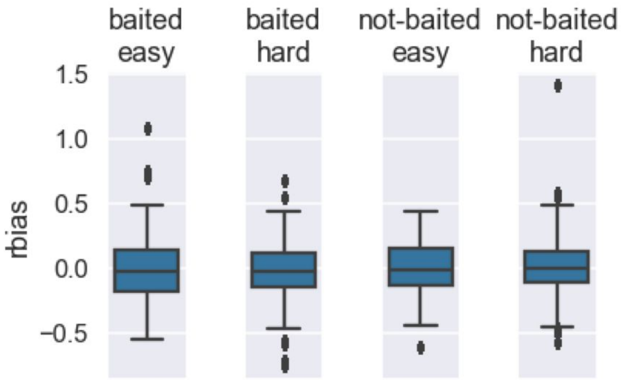
### Test Development: Getting to Reward

The main outcome of this test is bias towards the rewarded choice (here *more* or *less*), which is intended to be independent of discrimination performance. One of the challenges with this test has been to create a version that reliably produces bias in participants tested outside of the laboratory and across a range of device types. The original version used a face with either a short or long mouth, which differed slightly in length. This produced reliable bias towards the more frequently rewarded response (short or long) across multiple labs. In early versions of this task developed for TestMyBrain.org (2010 - 2012), we also found reliable bias with this format. The proliferation of device types since then, however, has made it much harder to control certain stimulus parameters (such as the difficulty of length discrimination) as the user can move the device and thus introduce considerable variation in viewing distance and angle.

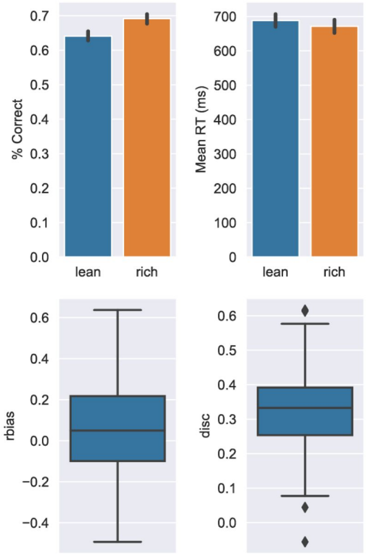
To address this, we moved to a different perceptual decision based on the same basic principle that we reasoned would be less dependent on viewing parameters: numerosity judgments. We have previously found these can be measured reliably across a range of devices. In this version of the task, participants were asked to decide if there were more blue or yellow dots on screen (see below), where one of the two responses was more likely to be rewarded.



We varied the difficulty of the task (easy, hard) and the way reward schedules were determined (baited, not baited) to create four versions. Unfortunately, *none of these versions* produced a reliable response bias (throughout this work, response bias and discriminability were computed using the formulas given in Pizzagalli et al., 2005). Boxplots of response bias in the first four versions of the task are shown below. These values did not differ significantly from zero by one-sample *t*-test in any case, all *p*s > 0.12.



The next iteration again involved numerosity judgments, but this time we used dots of a single color, provided a target (e.g., 24), and asked the participants to judge whether there were more or fewer dots than the target. This version was designed to be challenging so that reward contingencies (rather than perception) would drive behavior. Importantly, there was an equal number of trials with more versus fewer dots relative to the target. The participants, however, were rewarded three times more often for correct identifications of one "rich" stimulus type (e.g., more dots than the target) relative to the other "lean" stimulus (e.g., fewer dots than the target). Assignment of the stimulus types to the rich/lean conditions was counterbalanced.



The figure (left) shows accurate, RT, response bias, and discriminability in the new version of probabilistic reward test. As shown in the bottom left panel of the figure, with these changes we were finally able to elicit a reliable reward bias (one-sample *t*-test against zero, *t*(121) = 2.42, *p* = 0.017); note that data from 8 participants were excluded due to an excessive number of outlier response times (RTs). We also found the expected effect of stimulus type on accuracy. If the participant develops a bias to respond "rich", then his or her accuracy should be higher when the rich vs. lean stimulus is presented. Along these lines, the top left panel shows that accuracy (% correct) was significantly higher in response to the rich vs. lean stimulus, *t*(121) = 2.34, *p* = 0.039. RTs were also faster in response to the rich vs. lean stimulus (top right panel), although this effect was only marginally significant, *t*(121) = -1.86, *p* = 0.066. Finally, the bottom right panel shows that discriminability was significantly greater than zero, *t*(121) = 30.69, *p* < 0.001. Collectively, the data in Figure 4 show that this version of the task elicits the desired pattern of findings.

Psychometric data for this version of the task are given below.

**Test Details & Psychometric Characteristics**

This test includes 100 trials: 50 featuring the rich and lean stimuli, respectively. Rewards are only delivered when a stimulus is accurately categorized as “less” or “more” than the target, but correct identifications of the rich stimulus are three times more likely to be rewarded than correct identifications of the lean stimulus. Thus, the code is designed to deliver rewards on 30 (correct) rich but only 10 (correct) lean trials, ideally yielding a rich/lean reward ratio of 3:1. In this initial test, as in most studies using this task, we excluded as outliers trials in which the raw RT was less than 150 ms or greater than 2,500 ms, or in which the remaining, log[DP1] -transformed RT exceeded the participant’s mean±3S.Ds. Dropping such trials causes departures from the ideal. Nevertheless, in this initial test subjects earned 36 rewards on average, with a rich/lean reward ratio of 2.8 (26.7 rich rewards/9.58 lean rewards). These data indicate that the reinforcement contingencies experienced by participants were very close to what was intended, despite variations in behavior.

As noted above, response bias and discriminability are calculated using the equations given in Pizzagalli et al., 2005. The reliability of response bias scores is good (split-half reliability = 0.73; Spearman-Brown predicted reliability = 0.85), and they are normally distributed (Figure 5). There were no discernible relationships with age (Figure 6), sex (Figure 7), or education (Figure 8), but these null effects should be cautiously interpreted due to the small sample size.

**Appropriateness for Field Test Use**

*Device Effects.* We do not yet have enough data to evaluate specific device effects. *Participant Burden.* This test is burdensome to participants and has high attrition relative to other tests. Batteries containing this test had a mean participant rating of 3.5 out of 5 (compared to a site-wide mean of 3.7) and only 38% of people who began this test completed it (compared to 81% sitewide)

**Further Development**

Creation of a version of this test for field test use would depend on identifying better mechanisms of engagement. For example, this version used hypothetical incentives which are minimally rewarding. Appropriate use of the test might include true monetary (or other) incentives that can further elicit individual differences in reward learning and bias. Given the test’s reliability, the length could theoretically be reduced by 50% and still have acceptable reliability for bias scores (r > 0.7), which would help reduce participant burden.

Figure 13A. Distribution of Scores

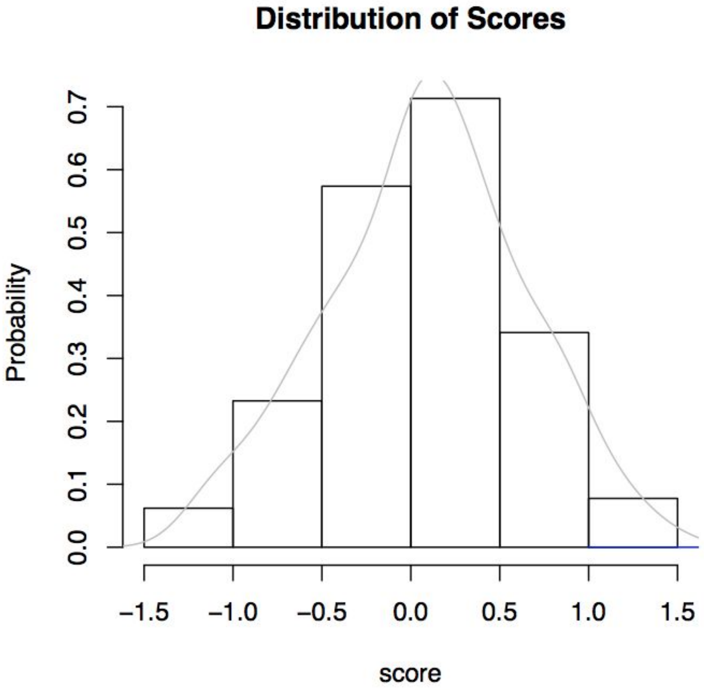


Figure 13B. Age-Related Differences in Performance (not significant)

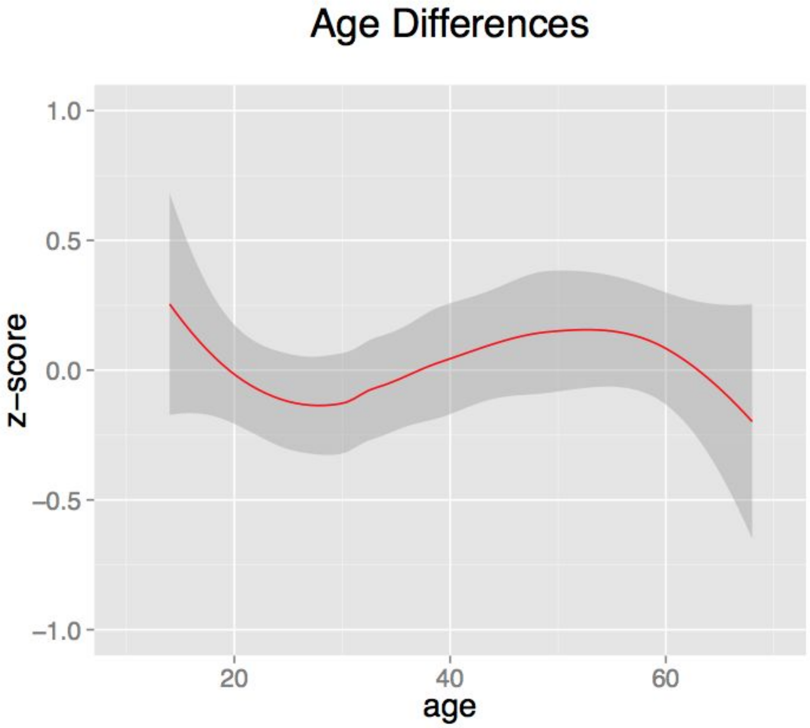


Figure 13C. Sex Differences in Performance (not significant)

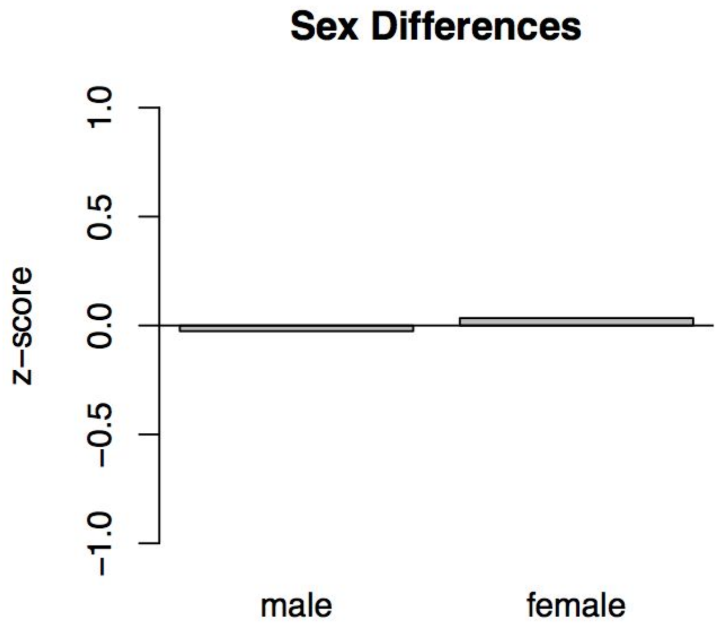
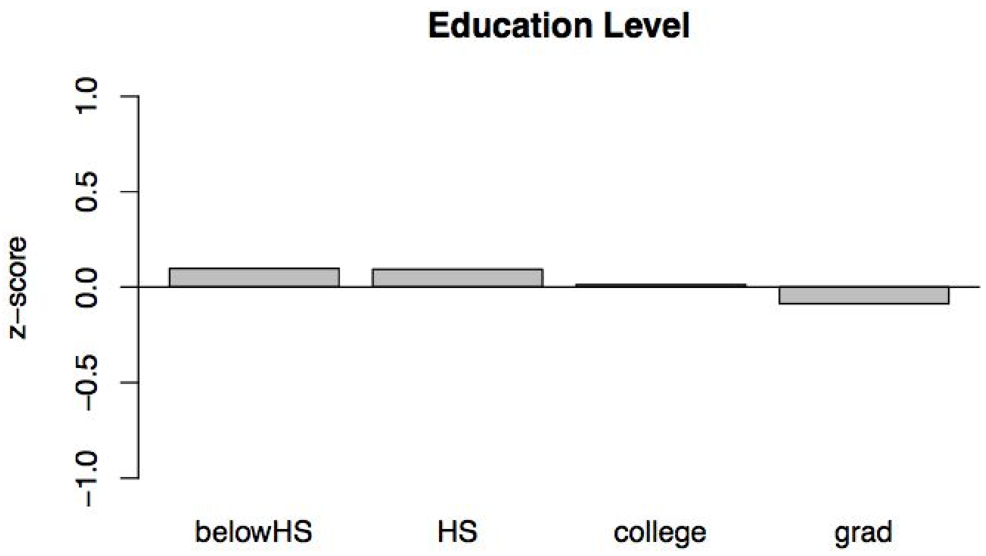
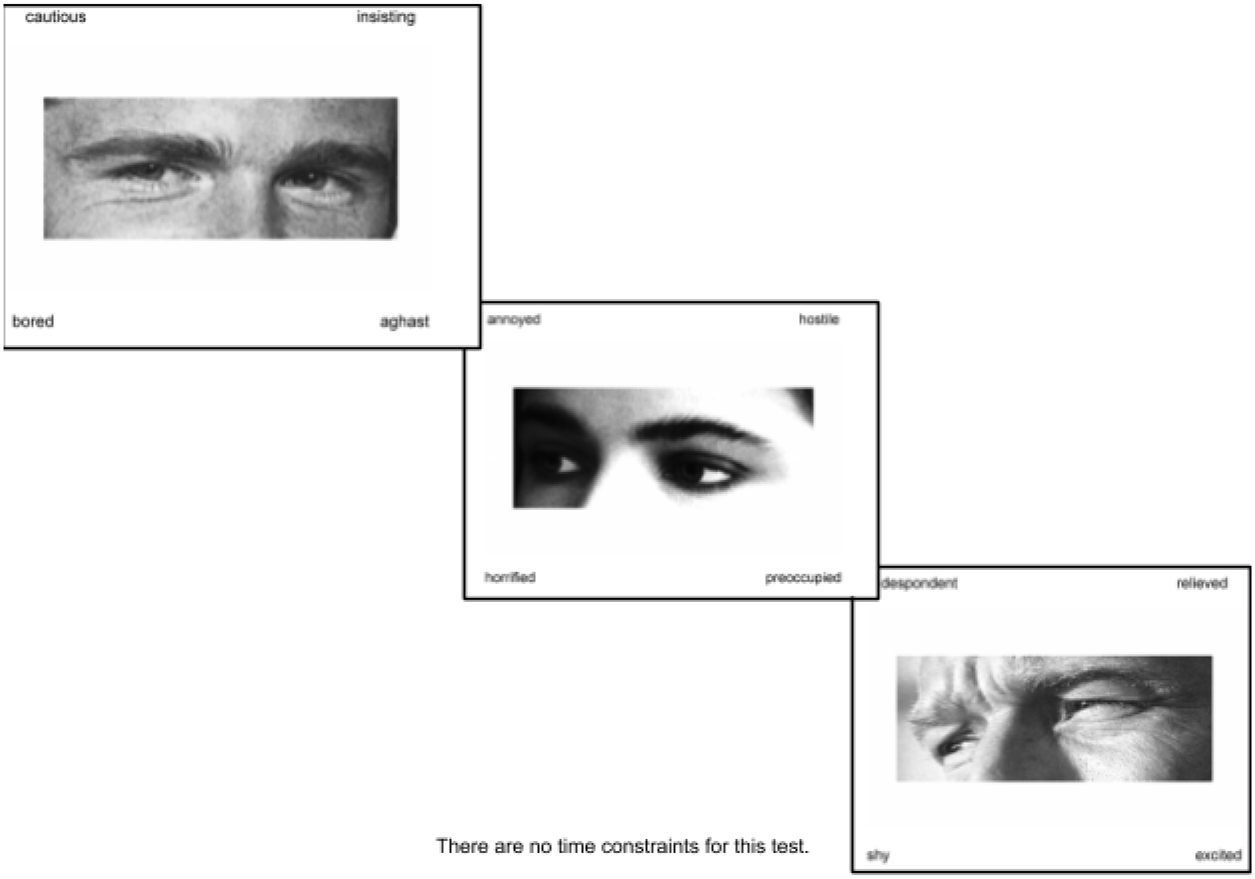


Figure 13D. Education-Related Differences in Performance (not significant)



### TMB Reading the Mind in the Eyes

Construct Measured: Social Communication / Reception of Facial Communication, Understanding Mental States
Duration: 10 minutes
Sample size from which normative data are available: 28,232
Demo Link: http://www.testmybrain.org/tests/mind_in_eyes/
Description of procedure: Decide which of four complex emotion words describes the mental state of someone based on just the eye region of their face.



There are no time constraints for this test.

This test is based on a well-validated and widely used measure of theory of mind or mental state inferencing (Baron-Cohen et al., 2001). Advantages of the task are that it is easily on a mobile device and performance can be interpreted with respect to a large body of existing literature. Drawbacks are that the test is long, with culturally homogeneous and potentially biased (with respect to race and sex) stimuli, as well as excessive reliance on complex vocabulary. The task is long but not viewed as particularly burdensome by participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. The Reading the

Mind in the Eyes test is included in the RDoC Council Workgroup Report on Behavioral Assessments so this task is designated **PRIORITY 1.**

### Current Applications

This test is virtually identical to the widely used format of the test developed by Baron-Cohen et al. (2001). Development of this specific format is currently included in the 23andme cognitive testing platform.

### Psychometric Characteristics

The primary outcome measure for the Reading the Mind in the Eyes test is accuracy, measured by proportion correct or number correct out of 37. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures would not be clear.

This test had good internal reliability (split-half) of 0.78, as estimated from a sample of 5000 participants who took this test on TestMyBrain.

Sociodemographic effects were estimated based on the scores of 18,492 participants for whom demographic data was available. This population has a mean age of 29.43 and is 55.94% female. Scores are relatively normally distributed, with minor ceiling effects (see Figure 14A). Performance is relatively consistent across adulthood: scores increase until approximately age 30 (with the sharpest increase during adolescence) and then are relatively stable throughout later adulthood (see Figure 14B). Female participants performed better, on average, than male participants (see Figure 14C). Performance increases with education, although there is no improvement in performance for participants with graduate-level education compared to those with only a college education (see Figure 14D).

One notable and potentially problematic characteristic of this test is that performance is highly associated with both race and ethnicity, with nonhispanic whites outperforming all other groups, sometimes with large effects sizes (e.g. Cohen's d = 0.75 comparing nonhispanic black to nonhispanic white participants). The combination of substantial vocabulary demands and ethnic homogeneity of (caucasian) faces in this test likely contributes to such differences.

This test showed no evidence of practice effects. First-time participants had a mean score of 26.08, while repeat participants had a mean score of 24.86.

### Validation

Performance on the Reading the Mind in the Eyes task is correlated with performance on other tests of emotion recognition (correlations are corrected for age when age data is available). It correlates moderately to highly with the Queen Square Face Discrimination Test (r = 0.46, N = 1538, 95% CI [0.42, 0.50]) and Morphed Emotion Identification Test (r = 0.51, N = 921, 95% CI [0.47, 0.56]). It is also highly correlated with vocabulary (r = 0.50, N = 7026, 95% CI [0.48, 0.52]). We did not find scores were significantly correlated with depression symptoms based on the Beck Depression Inventory (r = 0.055, N = 362, 95% CI [-0.16, 0.048]).

## Appropriateness for Field Test Use

*Device Effects.* Scores on the Reading the Mind in the Eyes test differ slightly between users of different digital devices, with laptop and desktop computer users having the highest scores and smartphone users scoring the lowest (iPhone mean = 24.67, SD = 6.57, N = 45; iPad mean = 26.78, SD = 4.09, N = 45; Macintosh laptop/desktop mean = 27.81, SD = 4.43, N = 1864). These differences could be due to the difficulty of seeing the details of each photograph on a small screen, since larger screen size appears to be associated with better performance. However, mean scores for each device group were calculated without controlling for demographic differences between users of different digital devices. Further research may be needed to quantify and (if necessary) mitigate the effects of differences between devices.

*Participant Burden.* This test is relatively well-tolerated by participants, though it is not as engaging as some other tests and is longer than many other tests. The average participant rating for batteries containing this test is 3.69 out of 5, close to the site-wide mean rating of 3.7. 76.0% of participants who begin this test complete it. This is somewhat lower than the sitewide mean (81%) likely due to the length of the test.

## Further Development

The current version of this test relies on ethnically homogeneous Caucasian face and complex vocabulary words, so the test is not appropriate in cohorts that are diverse in terms, ethnicity, or education (Dodell-Feder, Ressler, & Germine, in press). Versions of the test that include multiracial faces are recommended for broader applications and simpler vocabulary terms than the standard version.

The test is also very long compared to other tests of similar constructs, making it potentially more burdensome when combined with other measures.

Figure 14A. Distribution of Scores

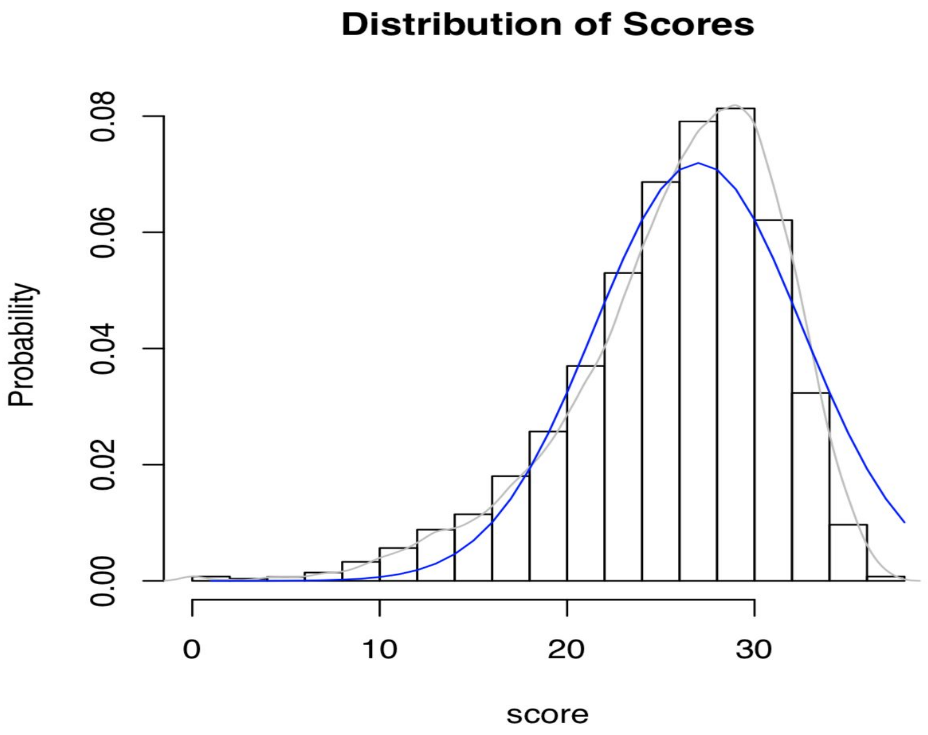


Figure 14B. Age-Related Differences in Performance

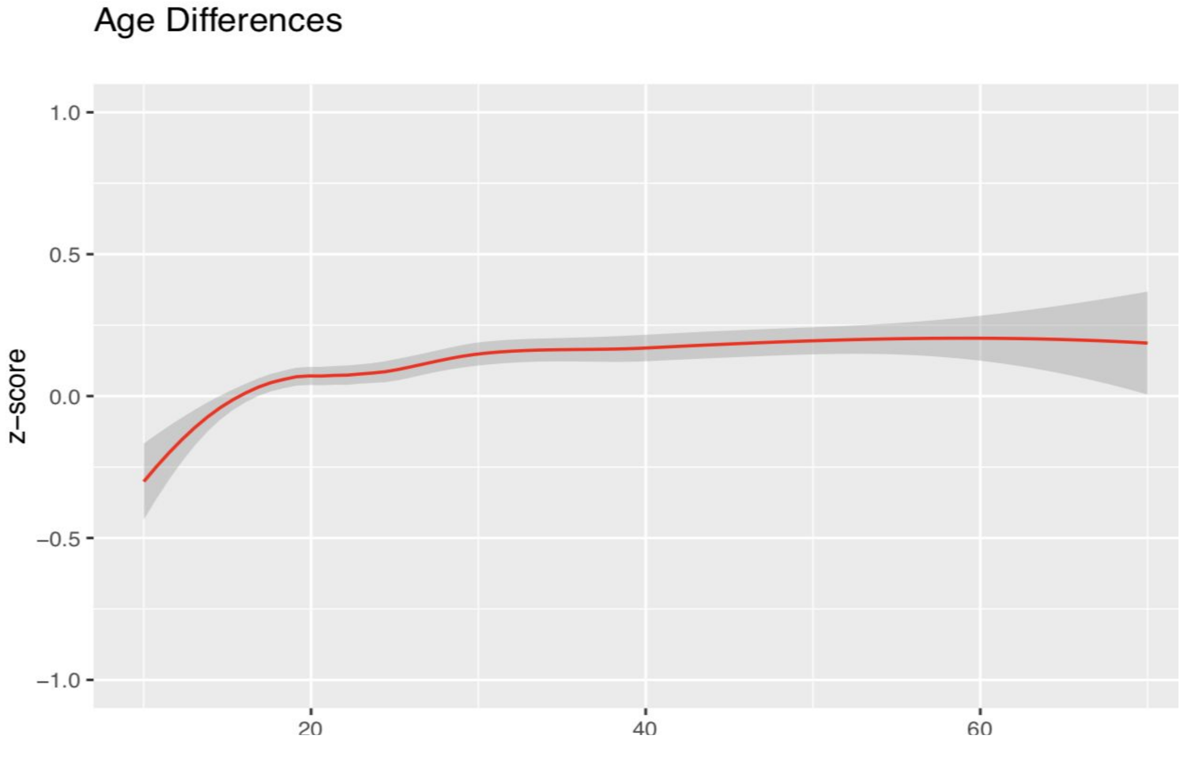


Figure 14C. Sex Differences in Performance

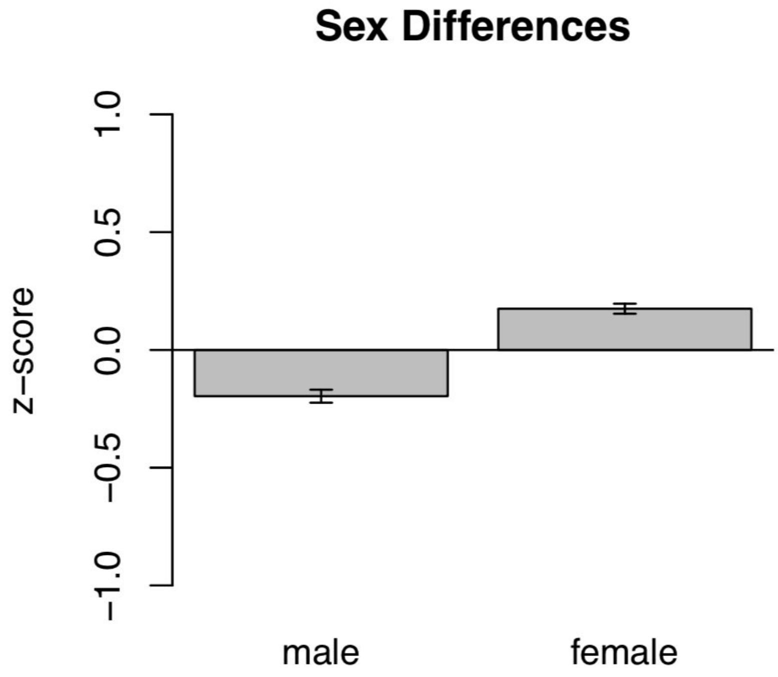
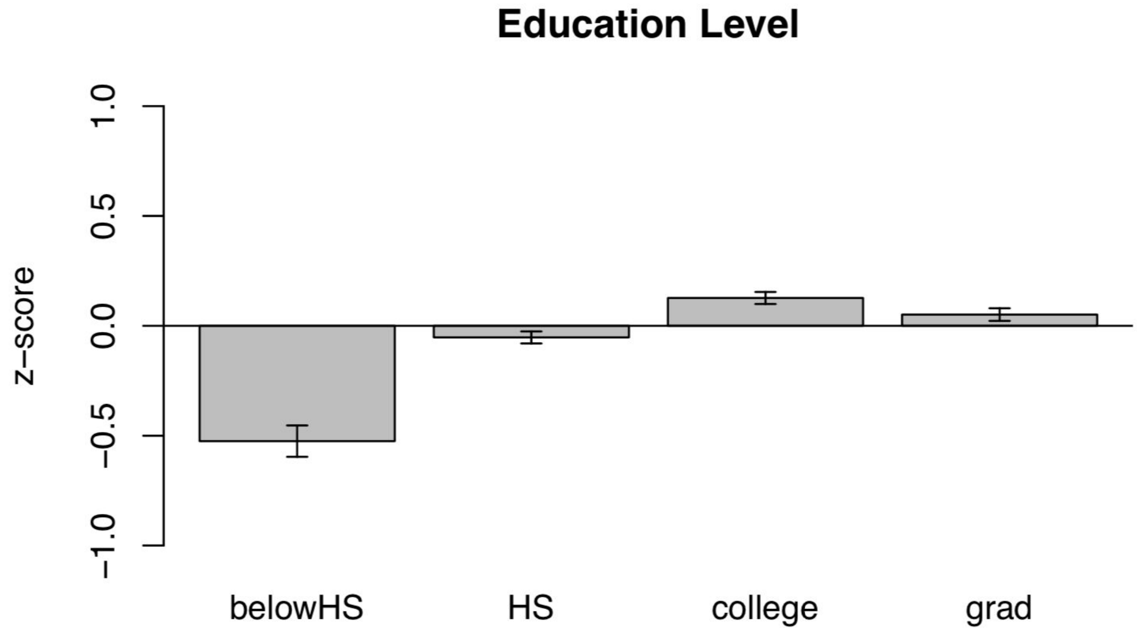
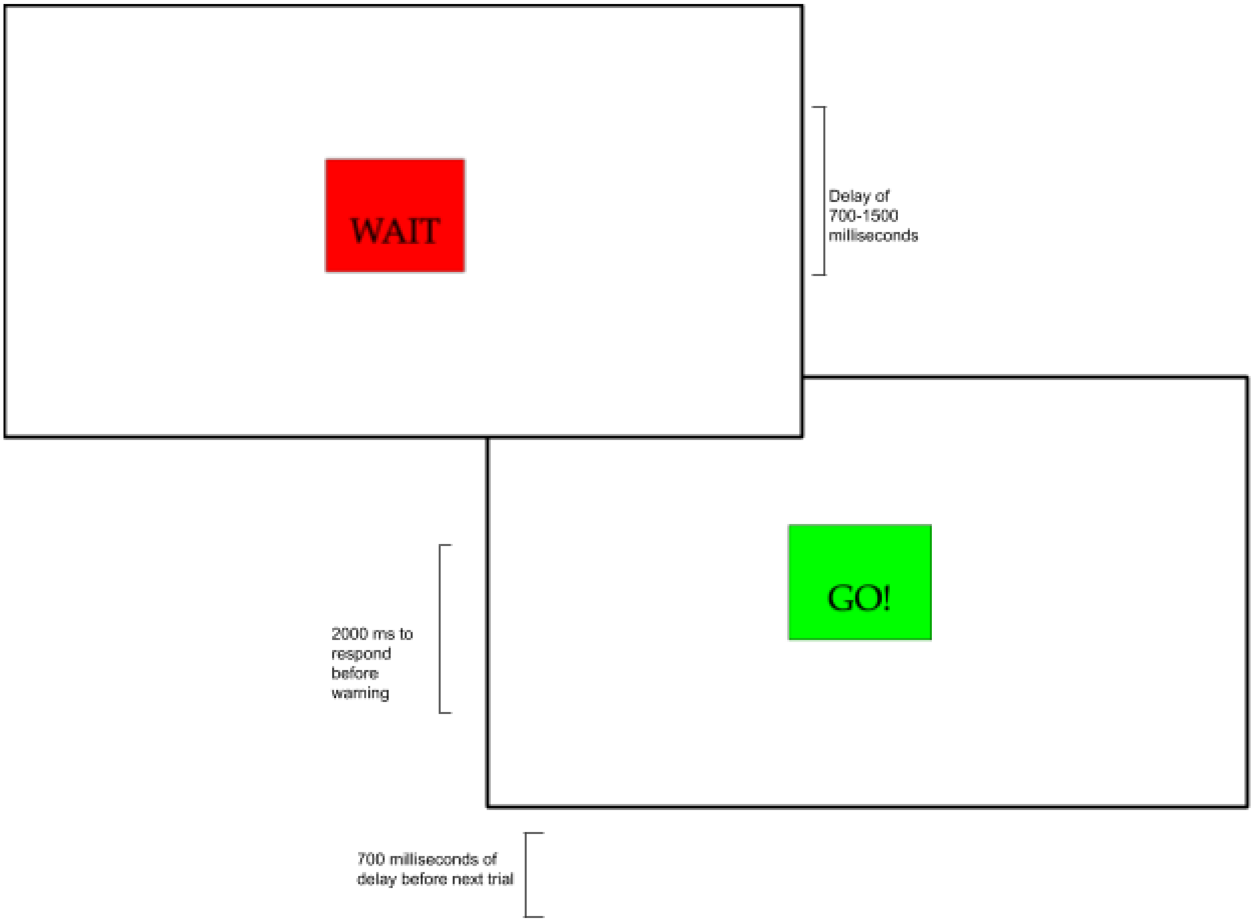


Figure 14D. Education-Related Differences in Performance



**TMB Simple Reaction Time**

Constructs Measured: psychomotor speed, response speed
Duration: 1.7 minutes
Sample size for which normative data are available: 49,001
Demo Link: http://www.testmybrain.org/tests/SimpleRTNew/SimpleRT.html
Description of procedure: Press a key whenever a green square appears.


WAIT
Delay of 700-1500 milliseconds
GO!
2000 ms to respond before warning
700 milliseconds of delay before next trial

This test is based a basic measure of simple reaction time, used to measure basic psychomotor speed where cognitive demands are minimized (Deary, Der, & Ford, 2001). Advantages of the task are that it is short, very sensitive, can be administered quickly and easily on a mobile device, and performance can be interpreted with respect to a large body of existing literature. Drawbacks are specific to a field test setting and include substantial device variance due to the interpretation of short and uncorrected response times. In combination with other tests (e.g. choice reaction time), however, this test can be used to better interpret cognitive performance on those tests.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Simple reaction time tasks are not included in the RDoC Council Workgroup Report on Behavioral Assessments, however, so this task is designated **PRIORITY 2.**

### Current Applications

The TMB Simple Reaction Time test is currently included in several major initiatives, including as part of the NIMH Aurora study and as part of the Broad Neuropsychiatric Phenotyping Initiative. Translation of the test into standard Chinese and Spanish is currently being funded by the Broad Institute.

### Psychometric Characteristics

The primary outcome measure returned by this test is the mean reaction time, which reflects a participant's ability to respond quickly to a stimulus (Lee & Chabris, 2013). This outcome is measured in milliseconds. To create a more standardized measure to present to participants, the mean reaction time can also be transformed into a score ranging from 0 to 100 (calculated as 10,000/mean reaction time), such that higher scores indicate faster mean reaction times.

The Simple Reaction Time test has excellent reliability; internal reliability (split-half) is 0.93, as calculated from a 5000-person sample of the participants who have completed the test on TestMyBrain.

Sociodemographic effects were estimated based on the scores of 47,024 participants for whom demographic information was available. This participant group had a mean age of 29.85 and was 45.68% female. Scores are normally distributed, with a small group of outliers with scores near the maximum value, likely due to participants who pressed the response button repeatedly as fast as possible rather than waiting for the cue stimulus (see Figure 15A). Performance is variable across the lifespan, with reaction times decreasing throughout adolescence before peaking at approximately age 20 and increasing throughout adulthood; this pattern is typical for reaction time-based tests (see Figure 15B). Male participants show slightly faster reaction times on this test than female participants (see Figure 15C). Effects of education on reaction time are minimal (see Figure 15D).

Practice effects on this test are minimal. First-time participants have a mean reaction time of 316.05, while repeat participants have a mean score of 305.53.

### Validation

Scores on this test shows moderate correlation with more complex tests of cognitive processing speed, such as reaction time in Choice Reaction Time (r = 0.40, n = 11178, 95% CI [0.38, 0.41]) and scores (inversely proportional to reaction time) and Digit Symbol Matching (r = 0.31, N = 21030, 95% CI [-0.32, -0.30]), and Trail-Making Test performance (r = 0.29, N = 8372, 95% CI [-0.31, -0.27]). This test is more modestly correlated with other measures that load on general cognitive ability, such as vocabulary (r = -0.13, N = 13455, 95% CI [-0.14, -0.11]). Thus, this test appears to both specifically measure cognitive processing speed and to reflect (to a lesser degree) broader cognitive abilities.

### Appropriateness for Field Test Use

*Device Effects.* Because the Simple Reaction Time test relies on measuring participant response times on a very brief scale, differences in device latency (the amount of time it takes for the device to register input) are likely to substantially affect scores. Among our participants on TestMyBrain, participants who used laptop or desktop had slightly faster reaction times than participants who used mobile devices (iPhone mean = 330.26, SD = 61.50, N = 5388; iPad mean = 345.23, SD = 70.90, N = 3152; Macintosh laptop/desktop mean = 298.22, SD = 57.81, N = 8175). Differences between laptop / desktop and tablet on this test are nearly a standard deviation in magnitude. Studies using this test must control for device type or use the same device type across participants.

*Participant Burden.* This task is well-tolerated by participants. The mean participant rating on TestMyBrain for batteries containing this test is 3.86 out of 5, compared to a site-wide mean rating of 3.7. 91.4% of participants who begin this test complete it.

## Further Development

The short reaction times in this test mean that the influence of device-related latency is substantial. While the test is appropriate for field test use, it must be interpreted with respect to other tests with similar latencies or device usage must be strictly controlled for scores on this test to be interpretable.

Figure 15A. Distribution of Scores

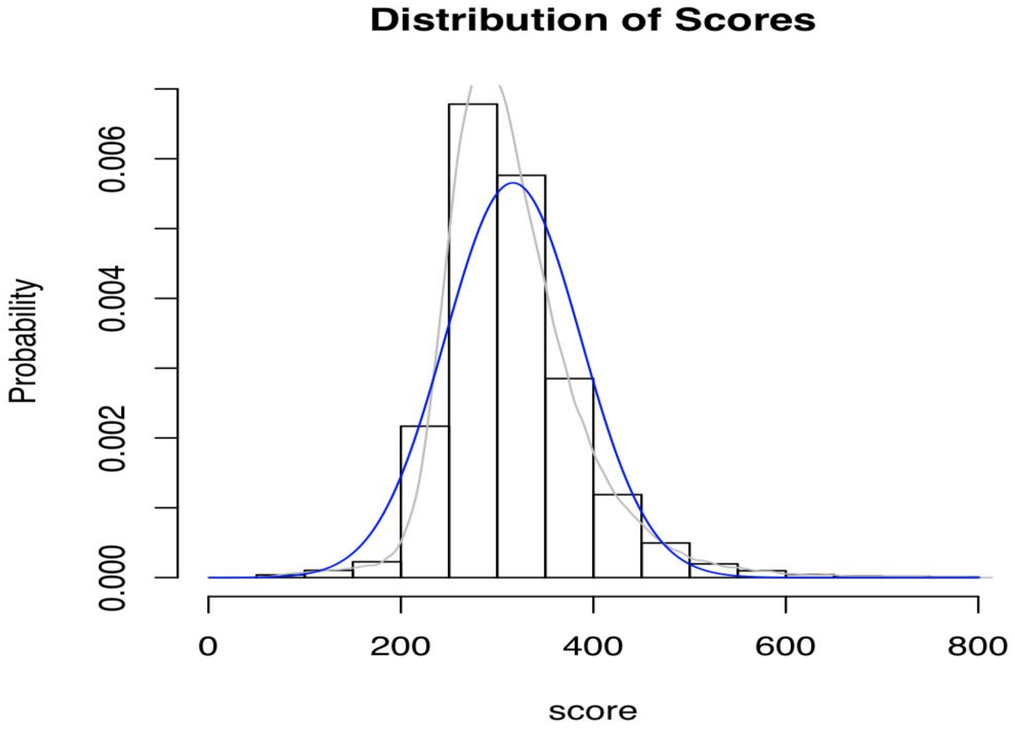


Figure 15B. Age-Related Differences in Performance

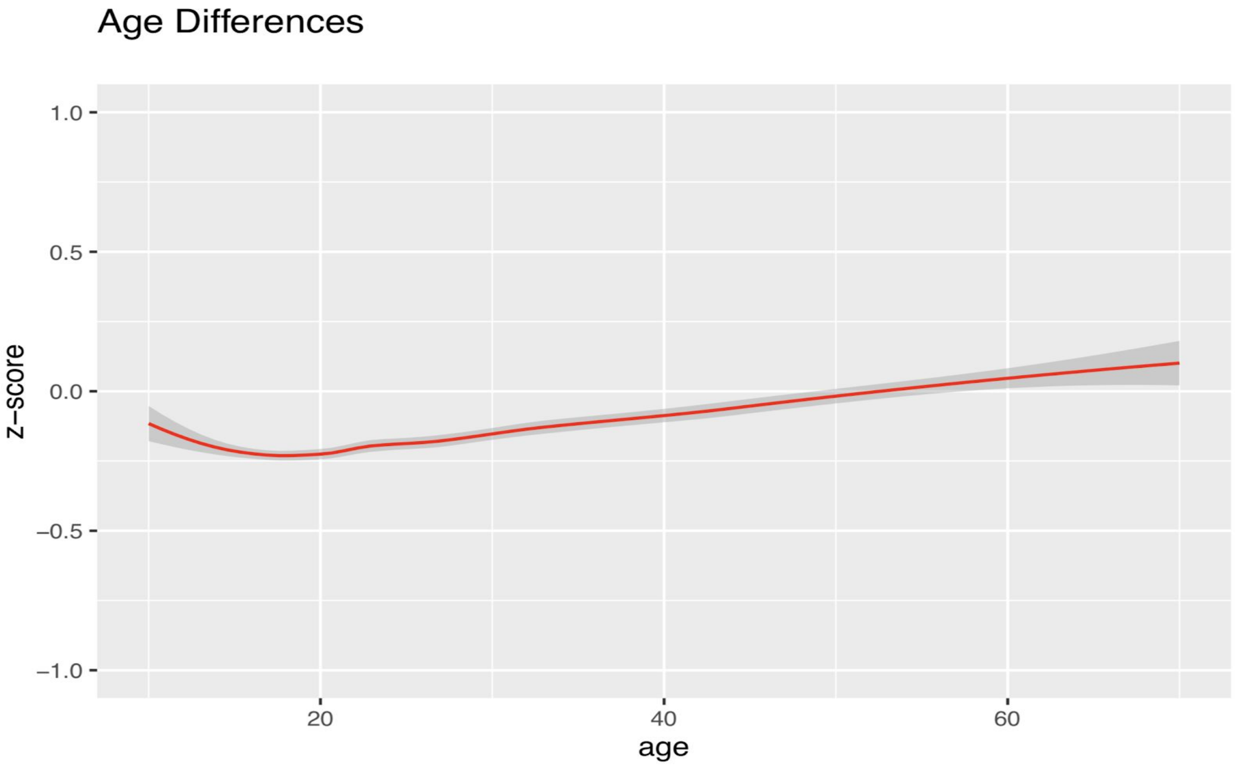


Figure 15C. Sex Differences in Performance

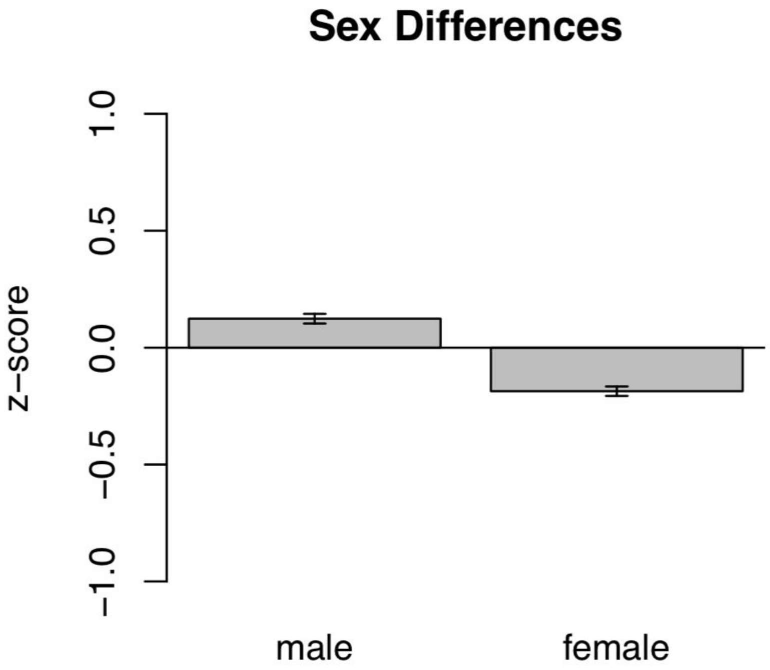
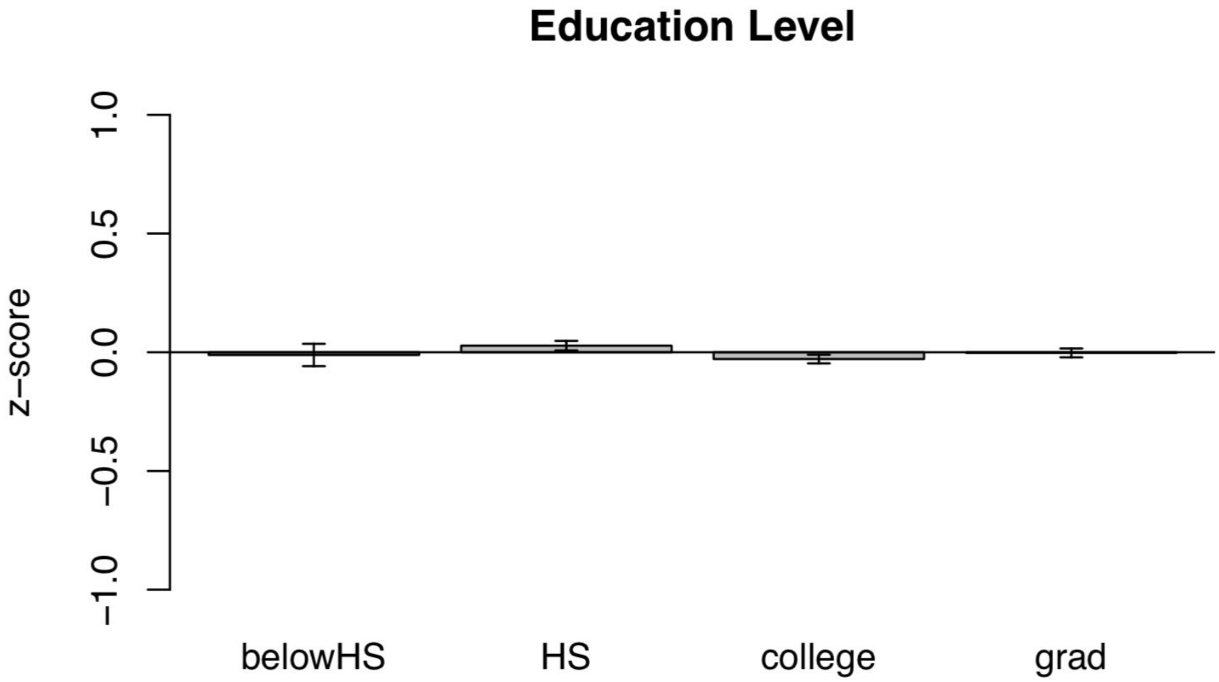
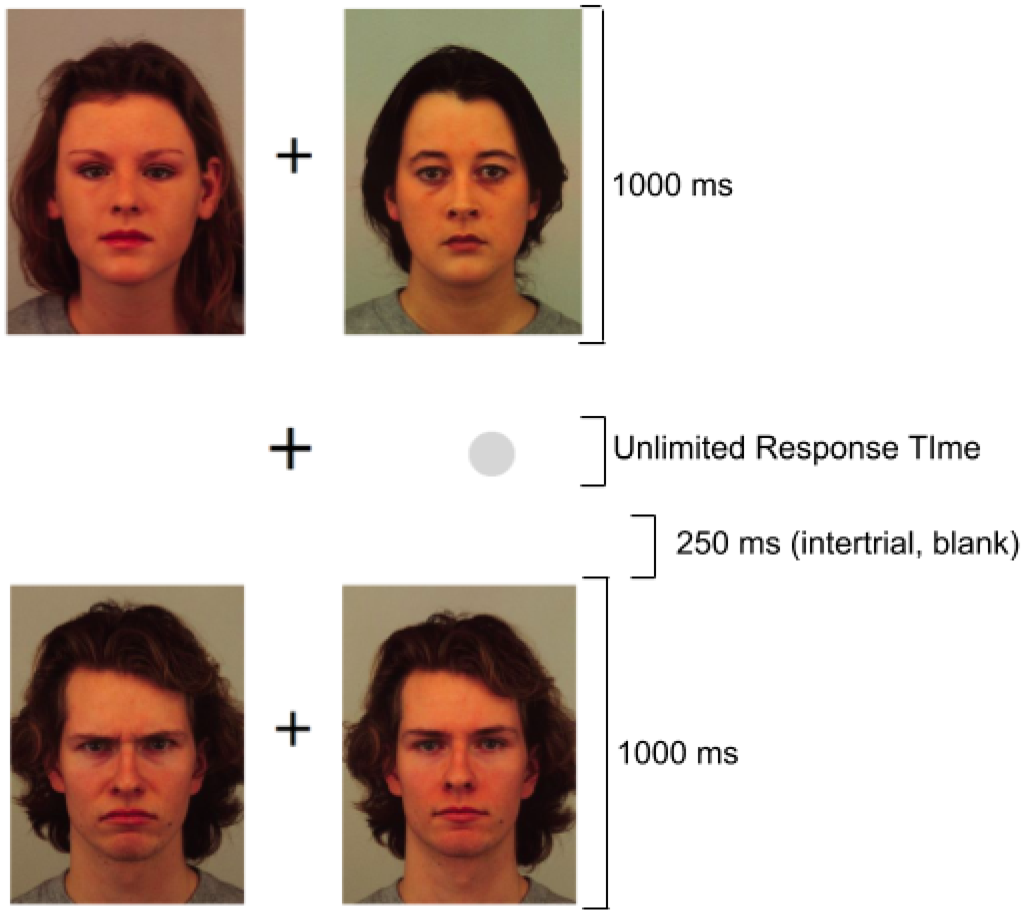


Figure 15D. Education-Related Differences in Performance



### TMB Threat/Neutral Dot Probe Test

Constructs Measured: potential threat, attention biases to emotion
Duration: 4.3 minutes
Sample size for which normative data are available: 1247
Demo Link: http://www.testmybrain.org/tests/dotprobe/dotprobe_ANGHAP_v7.html
Description of procedure: Judge whether a dot appears on the left or right side of the screen. Right before the dot appears, an angry, neutral, or happy face will appear in the same location as the dot.



This test is based on an adaptation of the Tel-Aviv University NIMH Threat / Neutral Dot Probe test for measuring attention biases to threat (Bar-Haim et al., 2010). Advantages of the task are that it can be administered easily on a mobile device. Drawbacks are that dot probe tasks have questionable reliability and validity for measuring between person variability. It is also questionable whether emotional faces will bias attention on the small screen of a mobile devices. The task is considered very burdensome to participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Spatial cuing tests are included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 1.**

### Current Applications

This test is being used as part of the NIMH Aurora Project, although it has been decided that this test will be dropped from future data collection efforts.

### Psychometric Characteristics

The Dot Probe task yields a measure of attention bias towards threatening stimuli. Typical scores range from -100 to 100. This score is calculated as the difference between a participant's median reaction time on trials when the dot is aligned with a neutral face and their median reaction time on trials when the dot is aligned with a threatening face (in trials where the set of faces consists of one threatening face and one neutral face).

The test is scored based on degree of bias towards or away from threatening faces. A score of 0 represents no bias, while a positive score indicates a bias towards threat (responding faster when the dot is aligned with a threatening face) and a negative score indicates a bias away from threat (responding faster when the dot is aligned with a neutral face).

This test has poor reliability for attention bias; analysis of all 1247 participants yielded an internal reliability (split-half) of .12 (not significant), while analysis of the 563 participants enrolled through the Aurora study showed a test-retest reliability of -.14 (not significant).

Sociodemographic effects were estimated based on attention bias for a sample of 274 participants for whom demographic information was available. This sample had a mean age of 34.9 and was 60.2% female. The distribution of scores is relatively normal, with a slight skew towards positive scores (indicating bias towards threat, although these differences are not reliable between individuals) (see Figure 16A). Performance is similar across ages (see Figure 16B). Age-residualized scores show that male participants show a slightly greater bias toward threat than female participants (see Figure 16C). There is no consistent trend in score by level of education (see Figure 16D).

### Appropriateness for Field Test Use

*Participant Burden.* The Threat/Neutral Dot Probe task is considered much less engaging than other tests. The average participant rating of batteries containing this test is 3.1 out of 5, compared to an average of 3.7 for all batteries hosted on TestMyBrain. 75% of participants who begin this test complete it, which is lower than the site average of 81%.

### Further Development

The remote / mobile form of this test does not have adequate reliable to justify its use, and any correlations with meaningful outcomes are likely to be spurious.

Figure 16A. Distribution of Scores

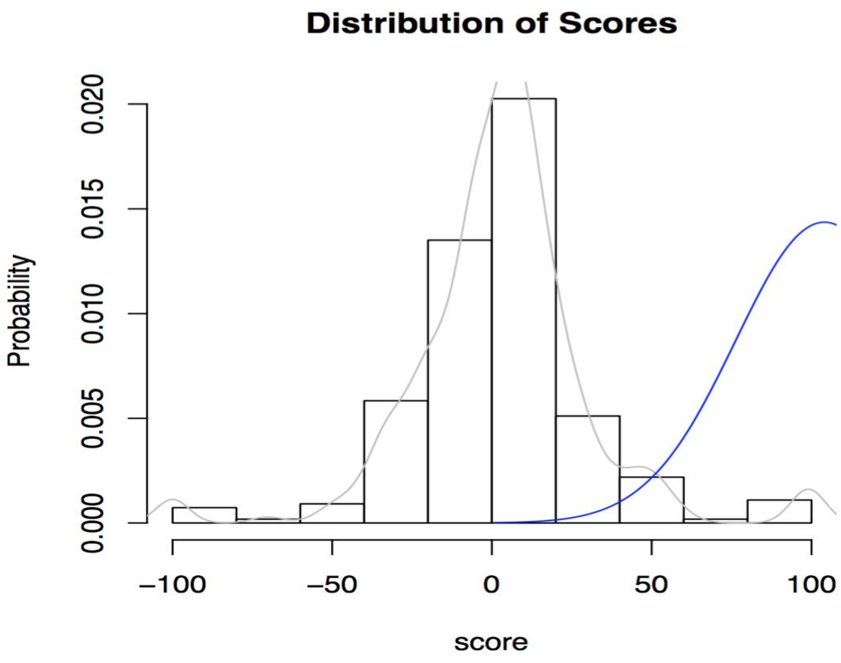


Figure 16B. Age-Related Differences in Performance

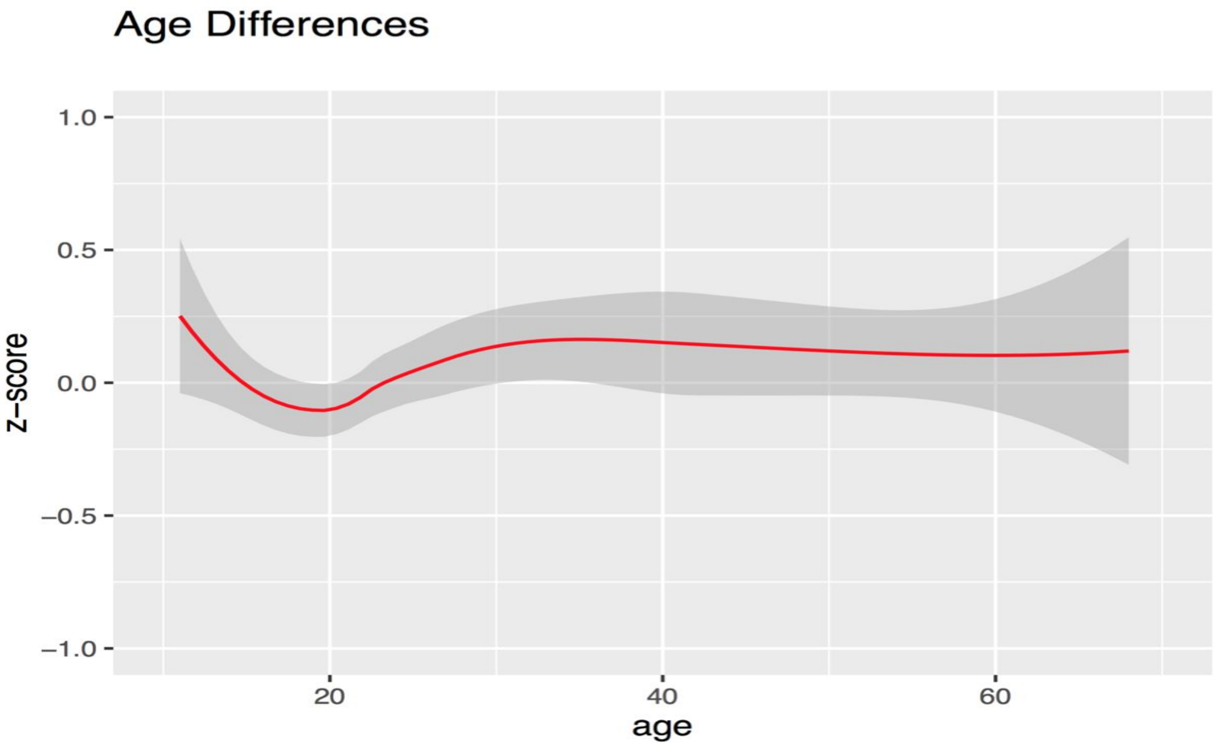


Figure 16C. Sex Differences in Performance

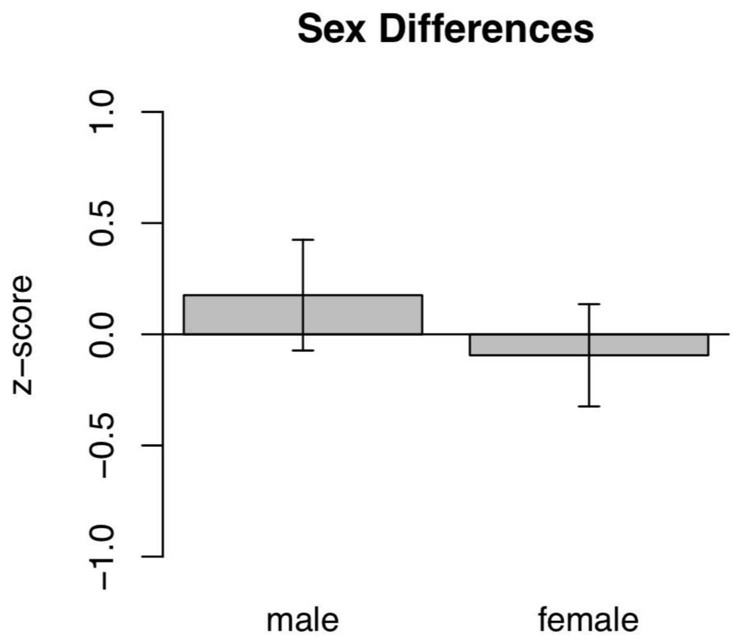
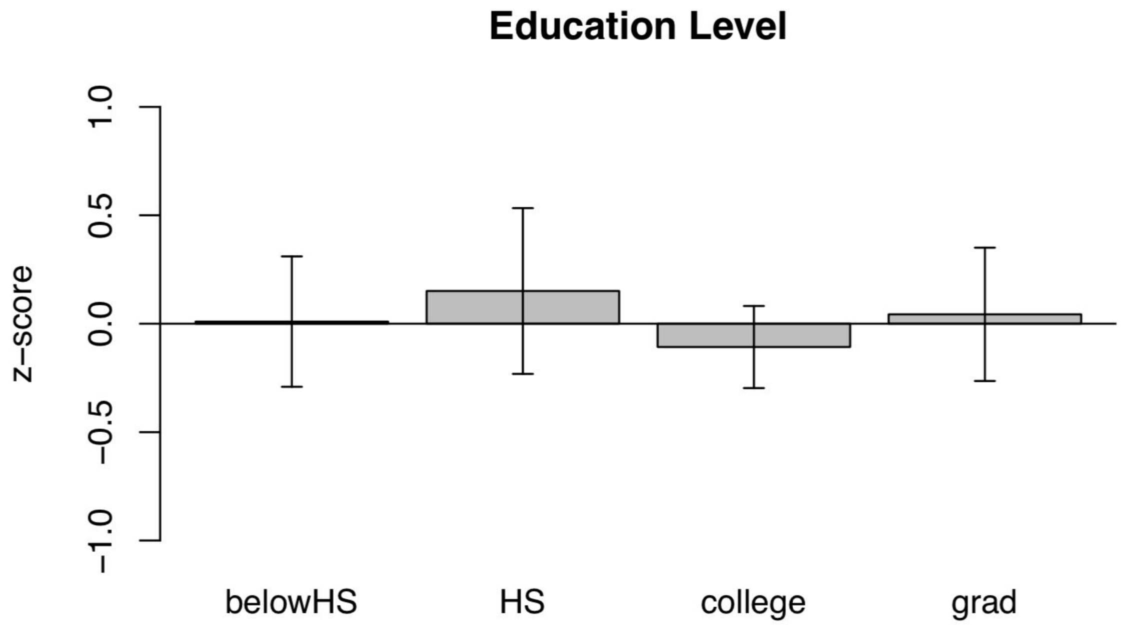
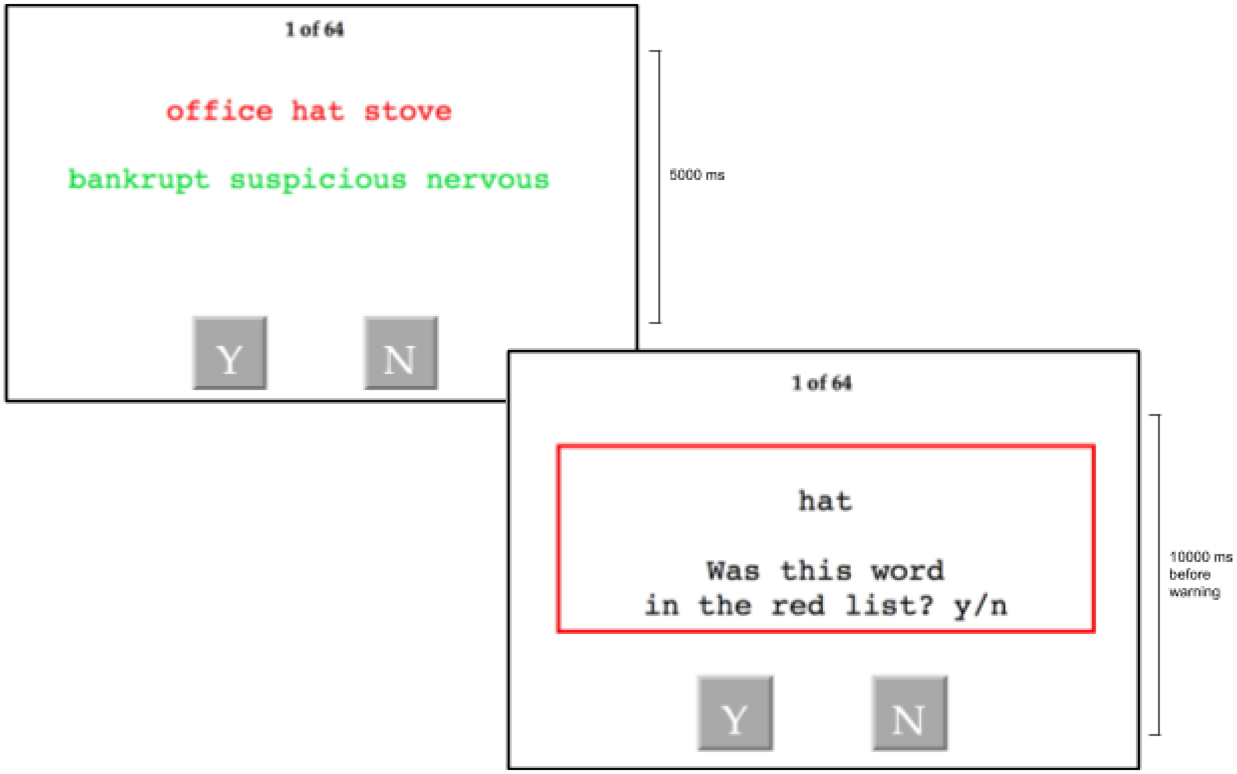


Figure 16D. Education-Related Differences in Performance



### Threat/Neutral Sternberg Memory

Constructs Measured: working memory, active maintenance, threat processing
Duration: 12.2 minutes
Sample size for which normative data are available: potentially available
Demo Link: http://www.testmybrain.org/tests/Sternberg/Sternberg.html
Description of procedure: Maintain a list of words in mind and decide if a target word was in the list. Words are neutral or threat-related and memory is cued by a colored square that matches the color of half the words immediately before test.



1 of 64

office hat stove

bankrupt suspicious nervous

1 of 64

hat

Was this word
in the red list? y/n

This test is based on an adaptation of the well-known Sternberg task format for measuring working memory and active maintenance, adapted to measure threat-related memory biases (Sternberg, 1975). Advantages of the task are that it can be administered easily on a mobile device. Drawbacks are that the test is lengthy due to the necessary inclusion of many conditions and extremely burdensome to participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Sternberg-type tests are included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 1.**

**Current Applications**

This test is currently being used in the NIMH Aurora study.

**Psychometric Characteristics**

The primary outcome measured by this test is threat bias in memory. Threat bias is calculated based on the difference between threat intrusion minus neutral intrusion where:
Neutral intrusion = median RT for correct *irrelevant neutral* trials (neutral target present in uncued list) - median RT for correct *new neutral* trials (neutral target not present)
Threat intrusion = median RT for correct *irrelevant threat* trials - median RT for correct *new threat*

This test can also be used to measure short-term memory (calculated as the proportion of trials in which the participant answered correctly) or reaction time, but the version presented in this report is specialized for the measurement of threat bias.

Threat bias as measured by this test shows relatively low reliability. Internal reliability (split-half) for threat bias was 0.14, while test-retest reliability (as calculated from the 997 participants enrolled through the Aurora project who completed this test multiple times) was 0.02.

Sociodemographic effects were calculated based on the 1373 participants for whom demographic data was available. This sample had a mean age of 30.96 and was 52.73% female. Threat bias scores are relatively normally distributed, with a large proportion of participants showing little to no threat bias (see Figure 17A. Threat bias is relatively consistent across the lifespan, but decreases slightly with age (see Figure 17B). Male and female participants show similar threat bias (see Figure 17C). Threat bias scores increase slightly with increased education, though this difference is not apparent between the most highly educated participant groups (see Figure 17D).

**Validation**

The Threat/Neutral Sternberg test is not significantly correlated with other tests of threat processing. This may be due to poor reliability. It does not correlate with threat bias in the Threat/Neutral Dot Probe test (r = -0.075, N = 324, 95% CI [-0.18, 0.034]), which measures differences in response speed between threatening and non-threatening cues.

**Appropriateness for Field Test Use**

*Device Effects.* Although threat bias does vary between users of different devices, these differences are not significant and are likely due to the noise of the measure (iPhone mean bias score = 74.08, SD = 618.22, N = 215; iPad mean bias score = -10.75, SD = 694.63, N = 68; Macintosh laptop/desktop mean score = 137.79, SD = 597.31, N = 119).

*Participant Burden.* This test is burdensome to participants and has high attrition relative to other tests. Batteries containing this test had a mean participant rating of 3.7 out of 5, close to the site-wide mean rating of 3.7, but only 61% of people who begin this test complete it (compared to 81% sitewide)

**Further Development**

It would be difficult to create a test based on this format that is ready for field test use. The test necessitates very long trials and yet the main threat bias effect is still relatively unreliable between individuals. This test is therefore not recommended for field test use.

Figure 17A. Distribution of Scores

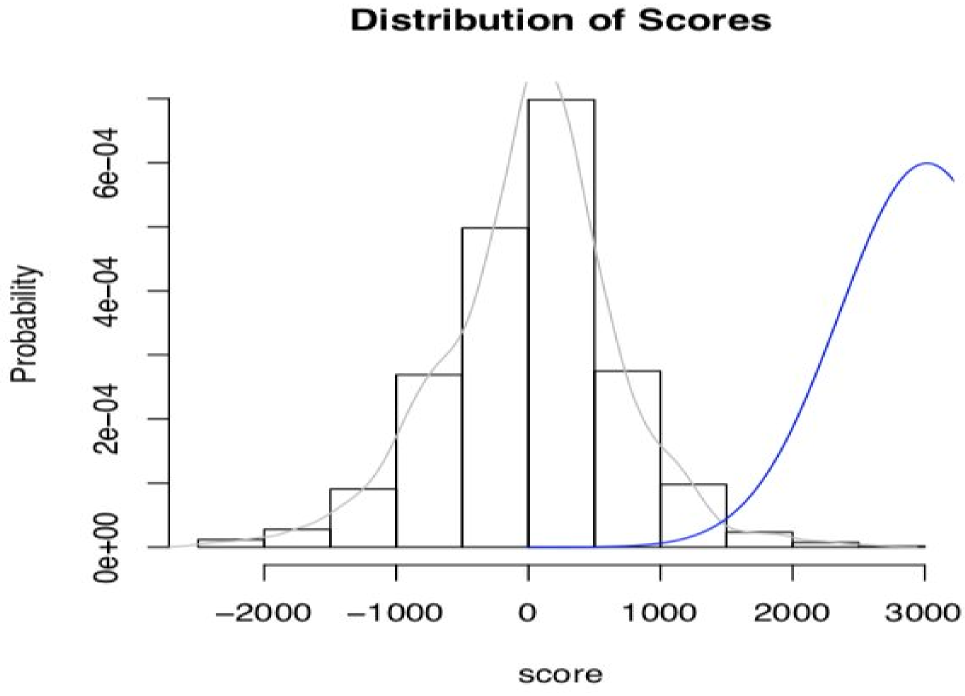


Figure 17B. Age-Related Differences in Performance

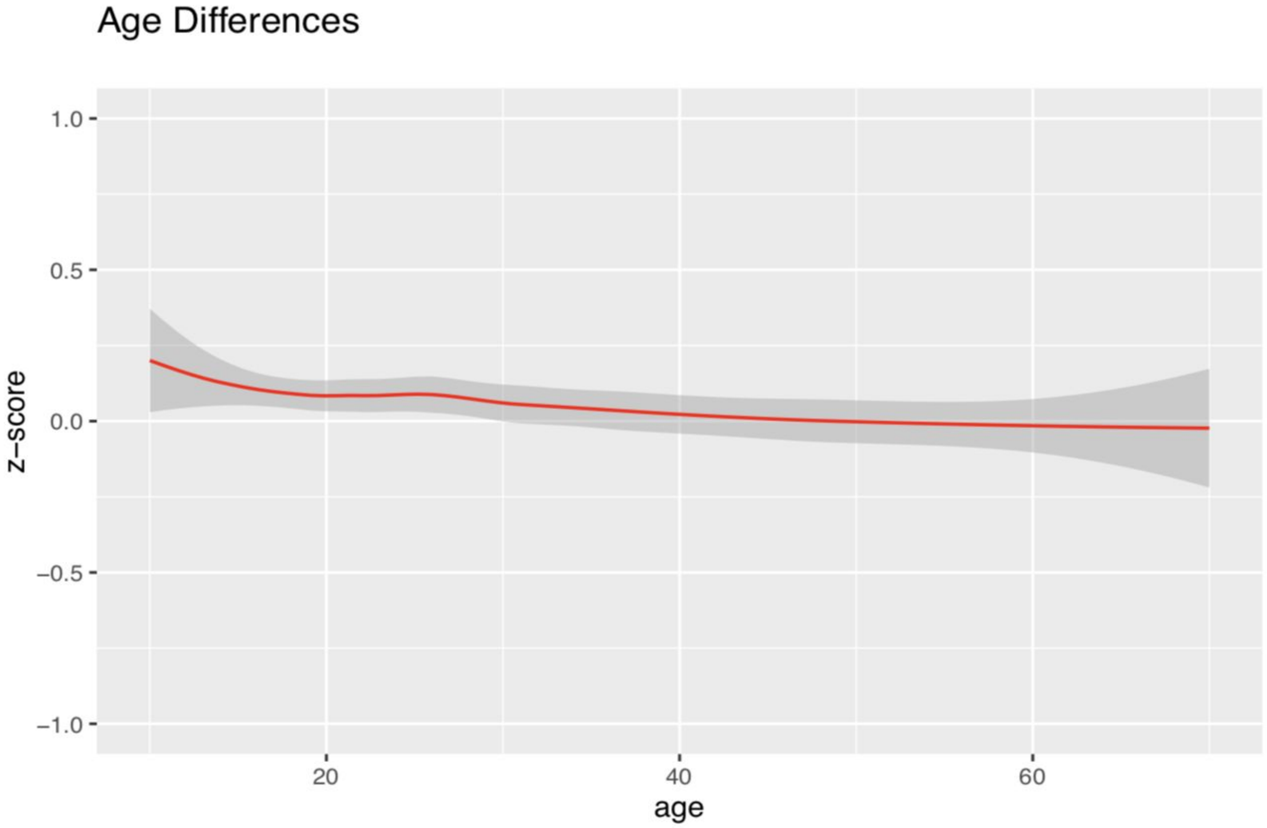


Figure 17C. Sex Differences in Performance

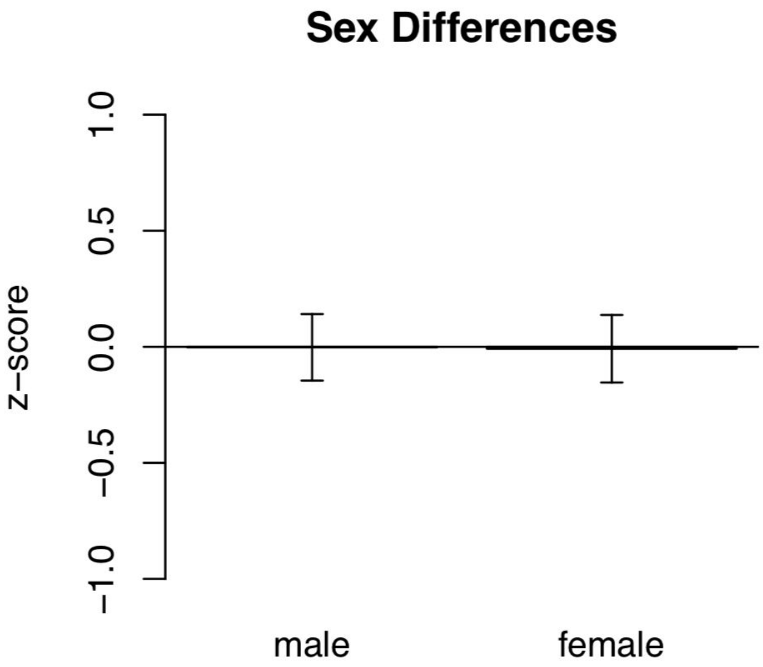
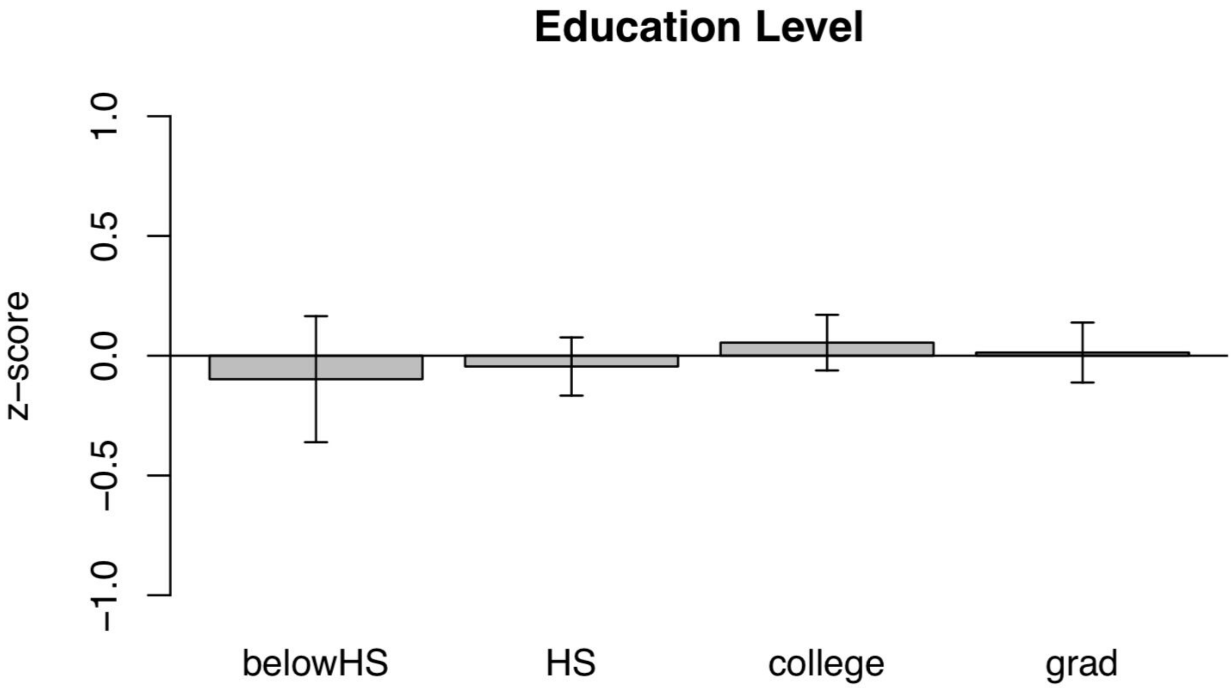
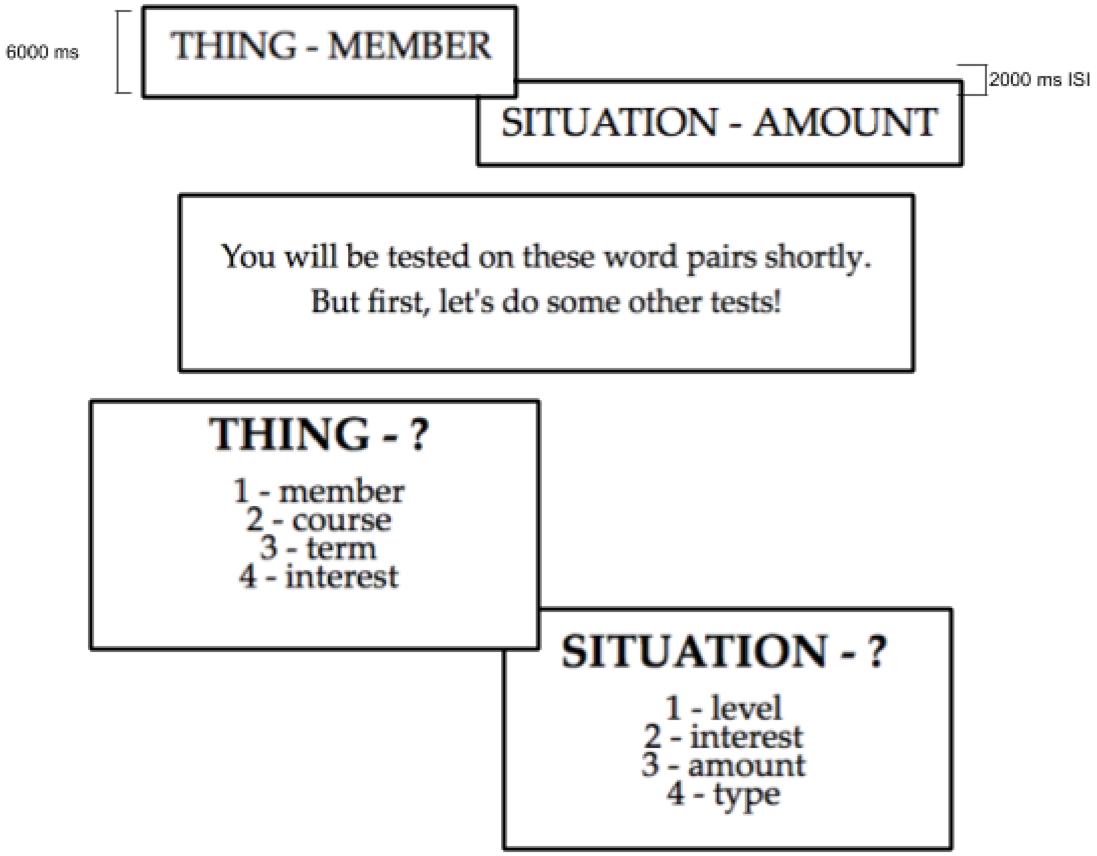


Figure 17D. Education-Related Differences in Performance



### TMB Verbal Paired Associates Test

Constructs Measured: Cognition: Declarative Memory, Language, Working Memory
Duration: 2.4 minutes memorization, 2.4 minuter test
Sample size for which normative data are available: 11,026
Demo Link: http://www.testmybrain.org/tests/vpa/mem2.html
http://www.testmybrain.org/tests/vpa/test2.html
Description of procedure: Learn and memorize a set of 25 word pairs. A subset of distractors repeat to increase difficulty and requiring learning of word pairs.



This test assesses verbal memory and episodic memory, and is adapted from standard paradigms for assessing context-specific encoding and verbal memory retrieval, as opposed to verbal recognition memory.  Advantages of the task are that it is short and can be administered quickly and easily on a mobile device.  The task is viewed as burdensome by participants.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test.  Episodic memory is not included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 2.**

## Current Applications

The TMB verbal paired associates test is currently included in several major initiatives, including as part of the 23andme cognitive platform, and as part of the NIMH Aurora study, as part of the Broad Neuropsychiatric Phenotyping Initiative. Translation of the test into standard Chinese and Spanish is currently being funded by the Broad Institute.

## Psychometric Characteristics

The Verbal Paired Associates test is scored based on the number of word pairs recalled correctly during the test phase out of 25. In order to more precisely target verbal and episodic memory (as opposed to visual memory), the version of the test shown in the demonstration link and pictures above uses only abstract words (e.g. “loyalty”, “satire”) that are difficult to visualize. Versions of this test that use more concrete words (e.g. “couch”, “tulip”) have also been designed; these versions of the task are easier, presumably due the ability to use visualization strategies for encoding, and therefore have ceiling effects, although such a distribution may be desirable for detecting impairment or in a general population sample.

Scores on this test are highly reliable, with internal reliability (split-half) for accuracy on the abstract (hard) version of this test of 0.82 based on a sample of 2073 participants. The concrete (easier) version of this has an internal reliability of 0.87, based on a sample of N = 399 in the Aurora study. Alternate forms test-retest based on the Aurora sample is 0.6.

Sociodemographic effects were estimated from a sample of 1980 participants for whom demographic information was available. This participant group had a mean age of 27.7 and was 52.5% female. The distribution of scores is skewed toward lower scores, consistent with the difficulty of this version of the task (see Figure 18A). Scores are relatively consistent across the lifespan, though there is a slight increase in performance throughout adolescence and a minor decline after age 60 (see Figure 18B). On average, female participants scored higher than male participants (see Figure 18C). Performance is correlated with education across all education levels (see Figure 18D).

As expected, people who take this test on multiple occasions show practice effects; first-time participants had a mean score of 12.61, while repeat participants had a mean score of 14.82 (Cohen’s d = 0.42). For this reason, any longitudinal study design should rely on alternate forms (as is currently being done in the Aurora project).

## Validation

Based on data from the Aurora project (concrete, easy version), performance on this test is modestly correlated with tests that rely on short-term memory such as forward digit span (r = 0.23, n = 494) and digit symbol matching (r = 0.36, n = 517). Performance on this test is also correlated with vocabulary performance (r = 0.37, n = 521). By contrast, the correlation is lower with tests of cognitive ability that do not involve significant challenges to memory or verbal ability, such as the Grad CPT sustained attention test (r = 0.066, n = 522) and simple reaction time (r = 0.17, n = 520). The version in Aurora also correlates moderately with Digit Symbol Coding (rho = 0.36, n = 517), a test that measures cognitive processing speed, visual processing, and visual memory.

### Appropriateness for Field Test Use

This task is easily adapted for field test use, but the length of the test may pose a barrier to completion. In order to control the time between learning and recall, this test should have another task interspersed between the learning phase and the test phase; this standardizes the conditions under which recall takes place, but may also increases participant burden.

*Device Effects:* The Verbal Paired Associates test can be administered across a wide variety of devices. Because this test is not scored based on reaction time or other time-dependent factors, differences between devices are unlikely to affect measured performance. Users of laptop or desktop computers score slightly higher than users of mobile devices (iPhone mean = 17.29, SD = 5.90, N = 1702; iPad mean = 17.63, SD = 6.05, N = 575; Macintosh laptop/desktop mean = 19.32, SD = 5.42, N = 1526). This is likely due to differences in demographic characteristics (such as age, sex, or education).

*Participant Burden:* The Verbal Paired Associates test is relatively well tolerated by participants, but the length of the test may be somewhat burdensome. Batteries hosted on TestMyBrain that contain a version of have an average rating of 3.8 out of 5, compared to a site-wide average of 3.7. Of the participants who reached the test portion of this task (after seeing all of the word pairs and completing an additional task between learning and recall), 97.1% completed it. However, only 43% of the participants who began the battery containing this test (which consisted of the learning phase, a distractor task to separate the two phases, and the recall phase) completed the entire battery. This suggests that the length of the three phases combined may pose a burden to participants.

### Further Development

The current version of this test would work well in a field test battery with minimal changes.  Alternate forms are already available for longitudinal designs (up to four assessments / four forms currently validated).  Concerns about attrition would need to be kept in mind, however, given the observed drop-off during encoding phases.

Figure 18A. Distribution of Scores

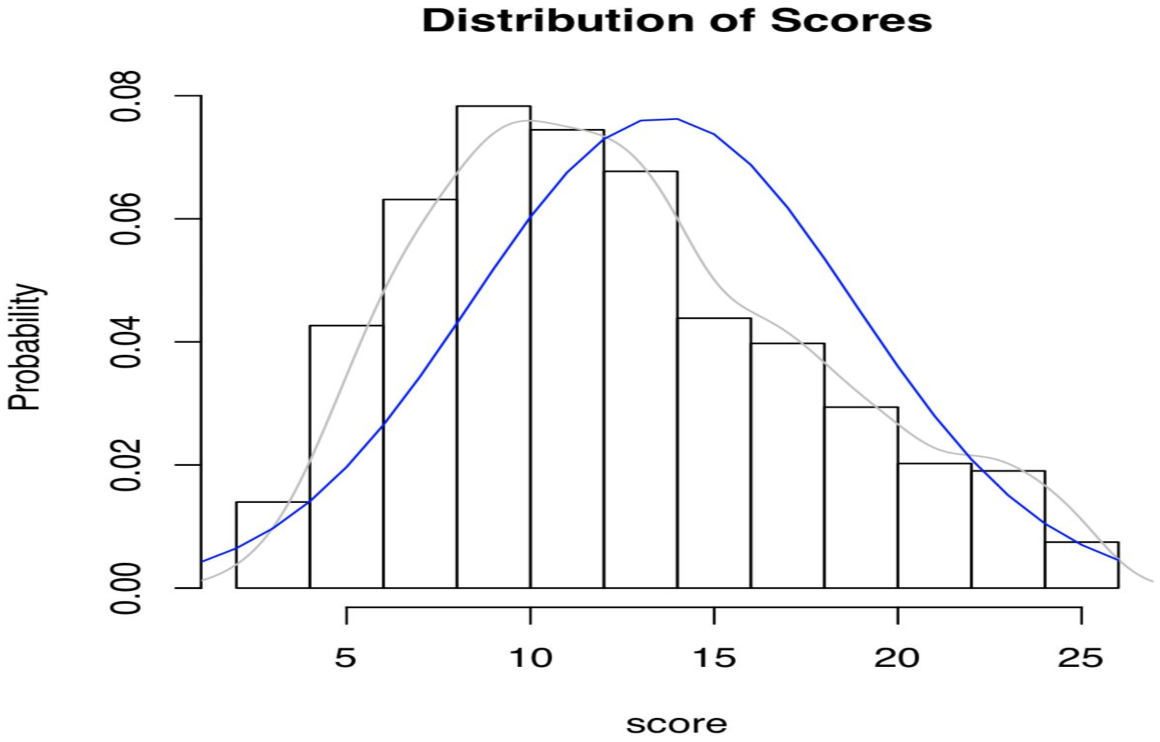


Figure 18B. Age-Related Differences in Performance

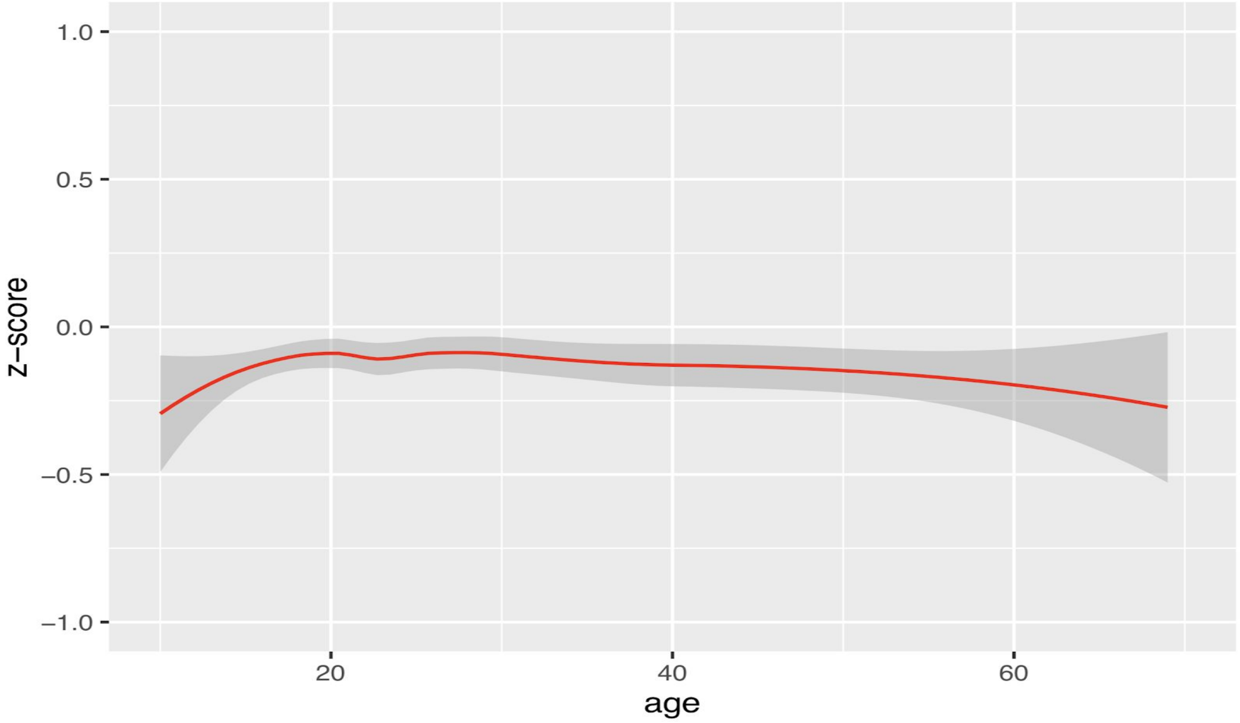


Figure 18C. Sex Differences in Performance

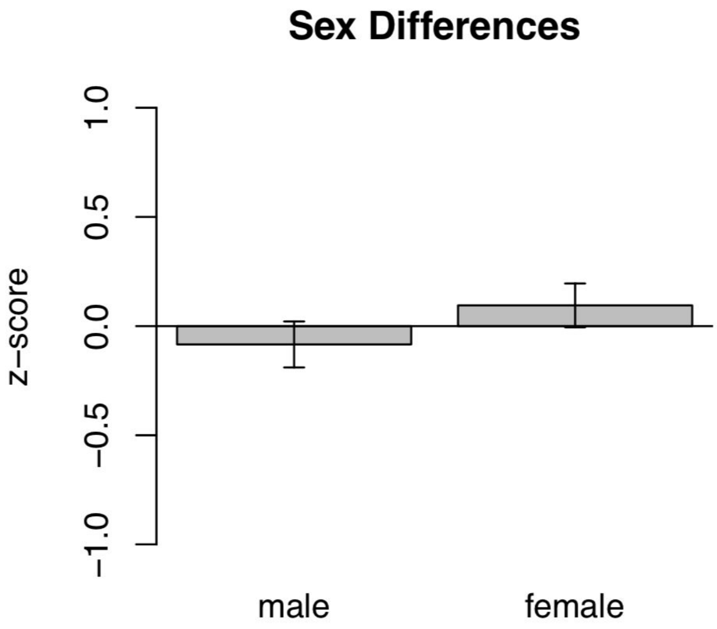
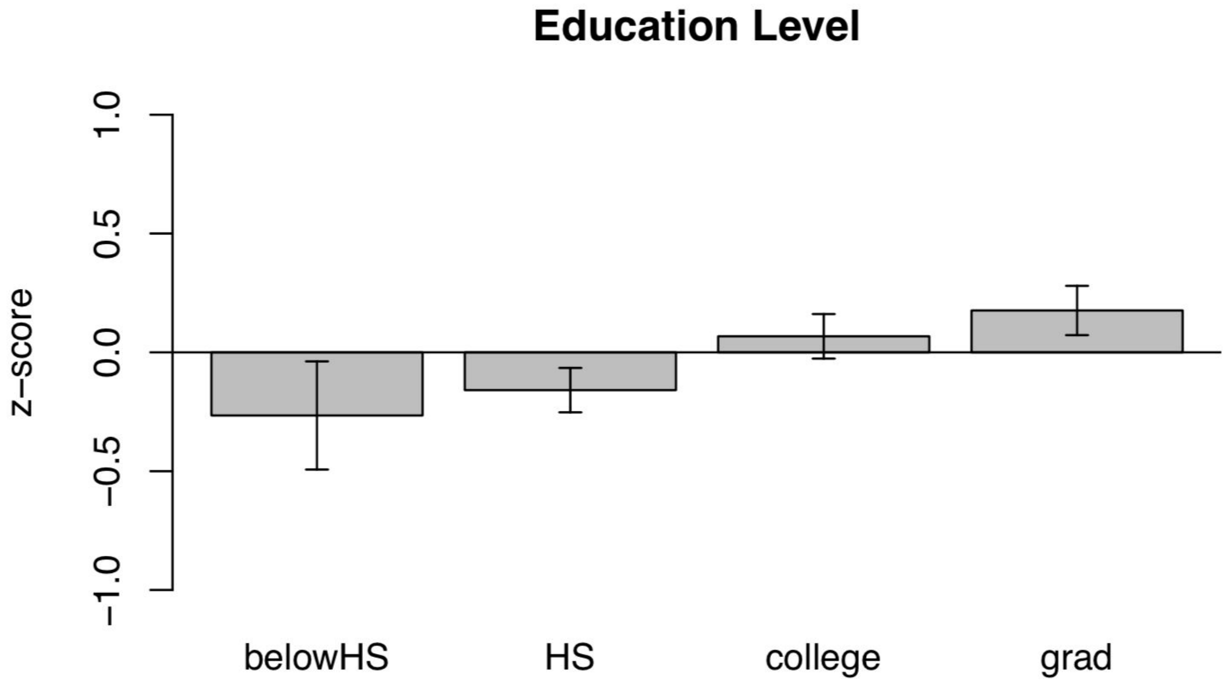
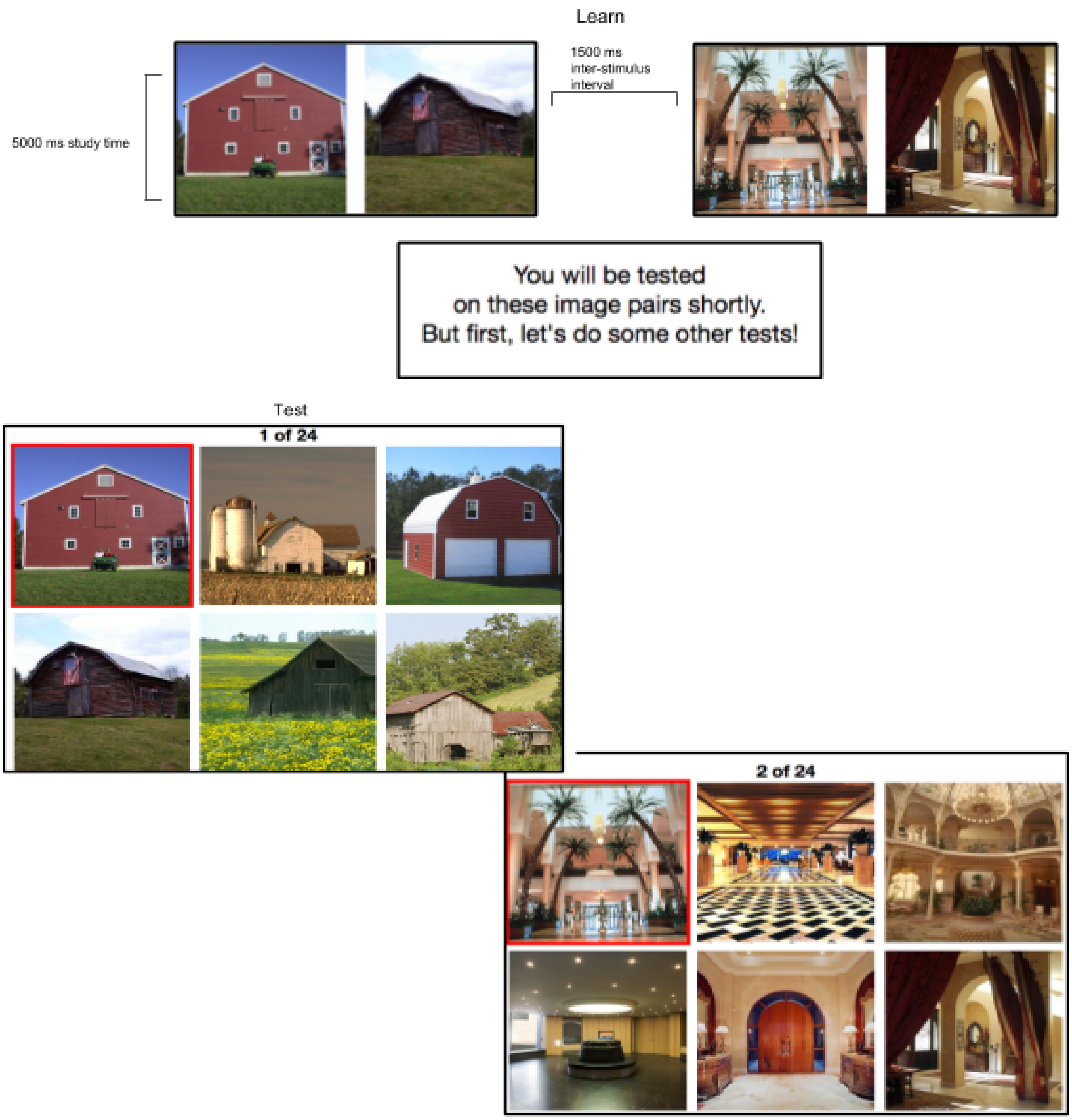


Figure 18D. Education-Related Differences in Performance



**TMB Visual Paired Associates Test**
Constructs Measured: Cognition: Perception, Declarative Memory, Working Memory
Duration: 2.4 minutes memorization, 2.4 minutes test
Sample size for which normative data are available: n=6,380
Demo Link: https://www.testmybrain.org/tests/visual_pair_assoc/VisualPAstudy_v2e.html
https://www.testmybrain.org/tests/visual_pair_assoc/VisualPAtest_v2e.html
Description of procedure: Learn and memorize a set of 25 image pairs. A subset of distractors repeat to increase difficulty and requiring learning of word pairs.



You will be tested
on these image pairs shortly.
But first, let's do some other tests!

This test assesses visual memory and episodic memory, and is adapted from standard paradigms for assessing context-specific encoding and memory retrieval, as opposed to visual recognition memory (Vellutino et al., 1975). Advantages of the task are that it is short and enjoyable and can be administered quickly and easily on a mobile device.

This test exists in a fully-implemented form for web/mobile self-administration on TestMyBrain.org and data are available that can be used to evaluate the test. Episodic memory is not included in the RDoC Council Workgroup Report on Behavioral Assessments, so this task is designated **PRIORITY 2.**

### Current Applications

The TMB Visual Paired Associates test is currently being further developed and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project as well as in the NIMH Aurora Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

The main outcome measure for this test is accuracy, in terms of proportion correct or number correct out of 24 trials. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures would not be clear.

The Visual Paired Associates test shows high reliability; the internal reliability (split-half) was 0.79, calculated from the 6380 participants who completed this test on TestMyBrain.

Sociodemographic effects were estimated based on scores for the 6014 participants for whom demographic data was available. This sample had a mean age of 34.11 and was 48.9% female. The distribution is normal, although with some ceiling effects (see Figure 19A). Performance is relatively consistent across the lifespan, but scores do increase slightly throughout adolescence and decrease after age 50 (see Figure 19B). Female participants have slightly higher mean scores than male participants (see Figure 19C). Performance increases with education (see Figure 19D).

Participants who take this test multiple times show slightly improved scores. First-time participants had a mean score of 15.74, while repeat participants had a mean score 16.48 (Cohen's d = 0.17).

### Validation

This test shows moderate correlation with vocabulary (r = 0.30, N = 1025, 95% CI [0.24, 0.35]), which also measures aspects of memory. It also correlates moderately with multiple object tracking, a test of visual perception and attention (r = 0.24, N = 5288, 95% CI [0.23, 0.28]).

### Appropriateness for Field Test Use

While this task is relatively simple, a practice round (in which participants choose the correct image with a prompt to guide them to it) is included to ensure that participants understand what is asked of them. Thus, difficulty understanding the task should not present a barrier to completion.

*Device Effects.* This test shows minor differences in performance between users of different devices; participants who took the test on a laptop or desktop computer had slightly higher scores than those who used mobile devices (iPhone mean = 15.18, SD = 4.30, N = 569; iPad mean = 15.56, SD = 4.53, N = 412; Macintosh laptop/desktop mean = 16.50, SD = 4.58, N = 1004). Because this test does not use timed outcomes to measure performance, differences in device latency are unlikely to impact scores on this test, but differences in screen size may affect participants' ability to see the images presented. Based on the comparison of iPad and iPhone, however, these differences appear to be minimal.

*Participant Burden.* This test is considered engaging by those participants that complete it - that is, ratings on the task are high, but attrition across batteries that include the test tend to have high attrition rates. The mean participant rating for batteries containing this test is 4 out of 5, compared to a sitewide average of 3.7. Of the participants who began the testing portion of this task, 98% completed it. However, only 56.2% of participants who began a battery containing this test completed the entire battery (which consists of a learning phase in which the pictures are presented, an intermediate task to control the time between learning and recall, and the test phase). For comparison, the sitewide battery completion rate is 75%. This suggests that participants are stopping during the learning phase or between the learning and recall phases, but those who begin the recall phase almost always complete it.

**Further Development**

The current version of this test would work well in a field test battery with minimal changes. Alternate forms would be useful, however, to enable use in longitudinal or pre/post designs and address expected practice effects over shorter time intervals.

Figure 19A. Distribution of Scores

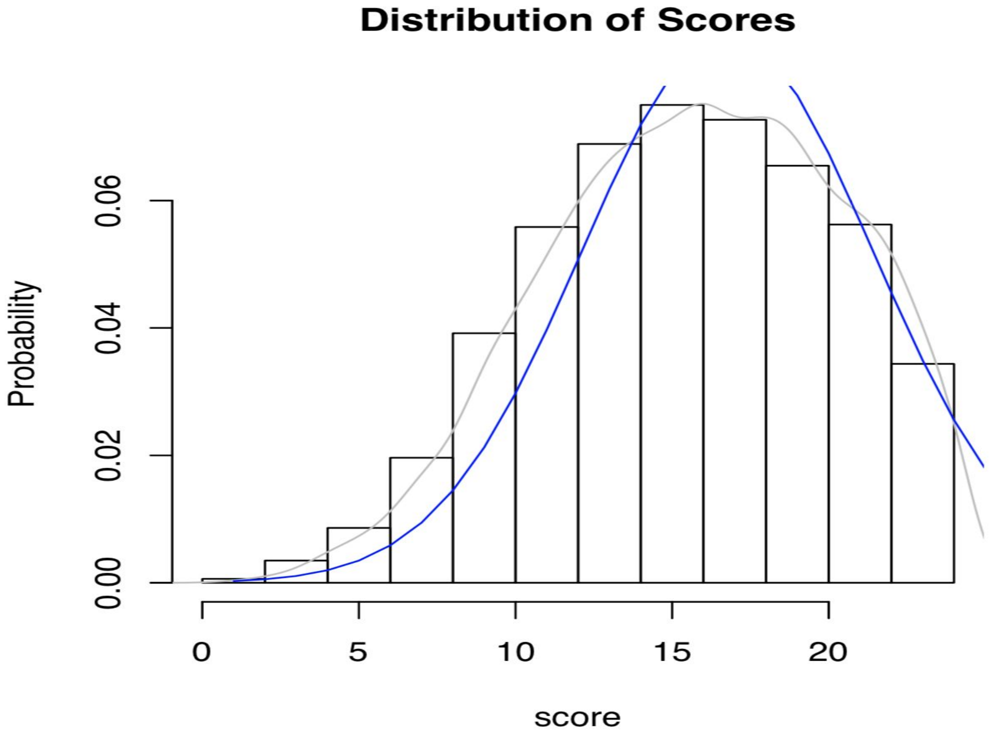


Figure 19B. Age-Related Differences in Performance

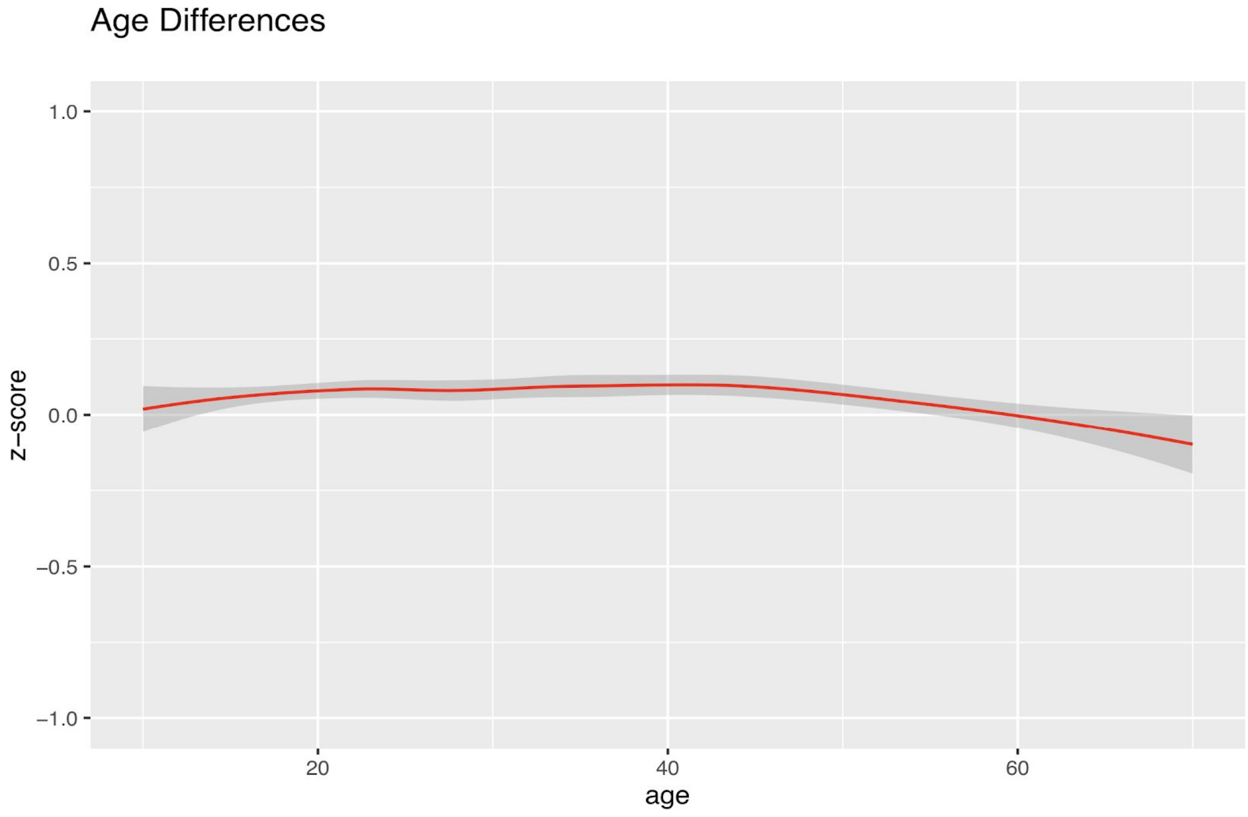


Figure 19C. Sex Differences in Performance

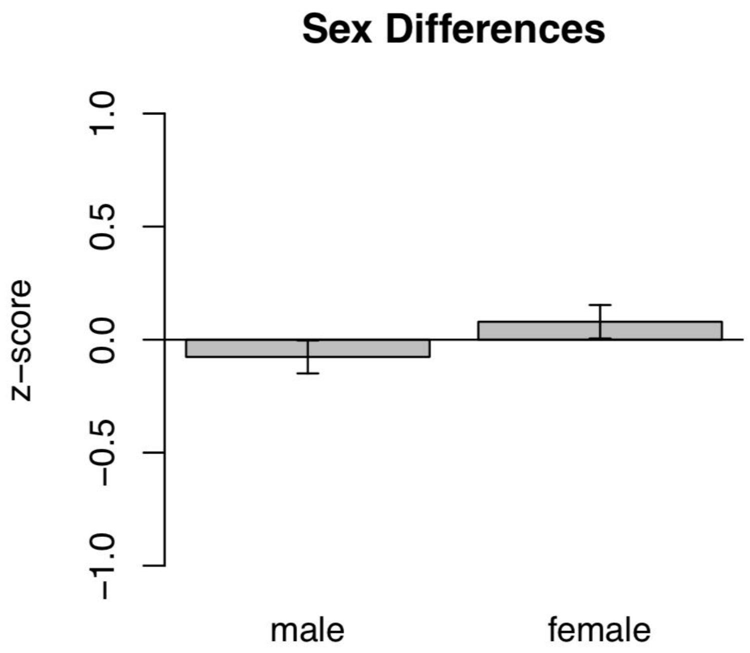
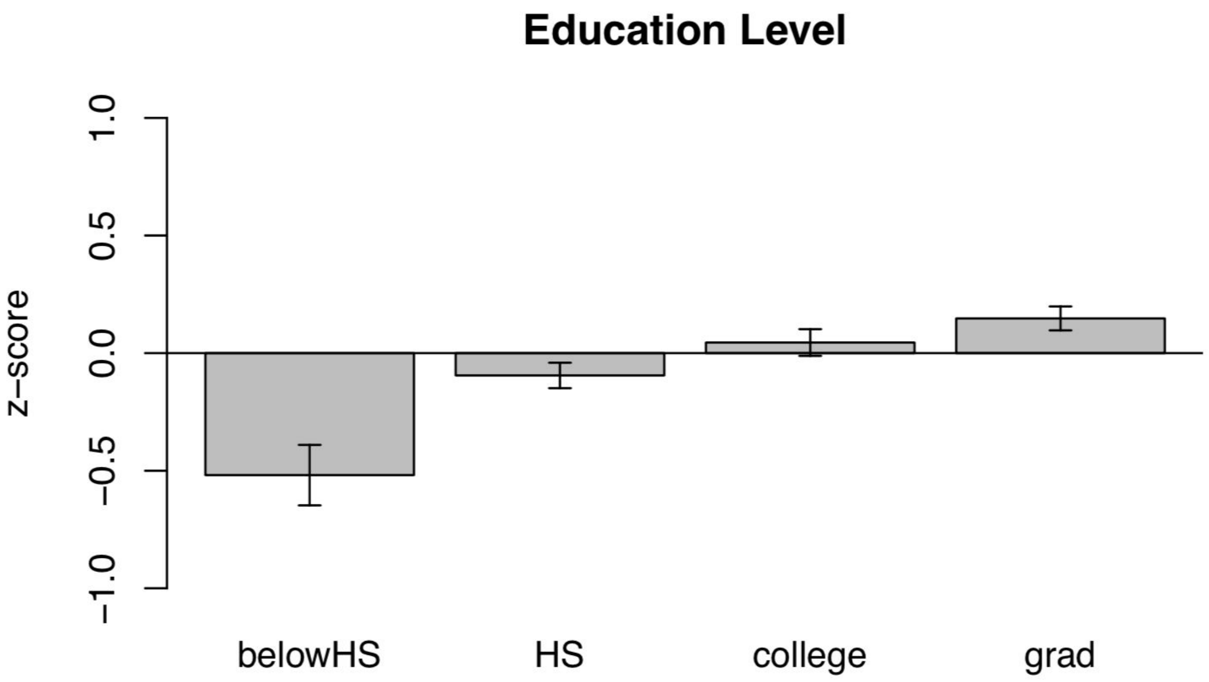
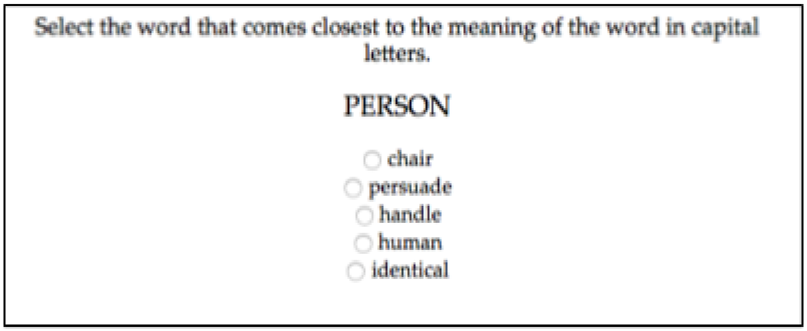


Figure 19D. Education-Related Differences in Performance



## TMB Vocabulary

Constructs Measured: Cognition: Declarative memory, language; Also general cognitive ability, general intelligence, crystallized intelligence, and verbal reasoning
Duration: 4 minutes
Sample size for which normative data are available: 40,169
Demo Link: http://www.testmybrain.org/tests/wordsum/index_v3e.html (easy example)
Priority: 1



Select the word that comes closest to the meaning of the word in capital letters.

PERSON

○ chair
○ persuade
○ handle
○ human
○ identical

No time constraints

The test consists of 20 test items (hard version) or 30 test items (easy version). For each item, participants are asked to select the word that is closest in meaning to the target word. Measures of vocabulary are among the best indices of verbal or crystallized intelligence, premorbid intelligence, and also of general intelligence more broadly. The TMB Vocabulary test was modeled after the well-validated Wordsum test used in the General Social Survey (Smith, Marsden & Hout, 2013).

### Current Applications

The TMB Vocabulary test is currently being further developed and evaluated as part of the Broad Institute Neuropsychiatric Phenotyping Project, by the 23andme personal genomics platform (adapted version), as well as in the NIMH Aurora Project. Translations are currently being prepared in standard Chinese and Spanish, funded by the Broad Institute.

### Psychometric Characteristics

Here we focus on accuracy (number correct or proportion correct) as the primary outcome measure or score. There are other reaction time-based measures that could be derived from this test (e.g. mean response time), but since this is not a speeded test the interpretation of these measures would not be clear.

The 20-item TMB Vocabulary test is twice the length of the 10-item Wordsum, and this produces the expected boost in reliability (Cronbach's alpha = 0.83 for Vocabulary in the present sample versus 0.68 for Wordsum; Cor et al, 2012).

Sociodemographic effects were estimated based on: hard 20-item version, N = 47,559 participants and easy 30-item version, N = 50,148 participants. The distribution of scores is relatively normal, with some ceiling effects, particularly on the easy version of the test and for older age groups (see Figure 20A and 20B). Performance is variable across the lifespan, with increases in performance across the full age range included in our analyses (see Figure 20B). Based on age residualized scores, there is a small gender difference that favors females (gender differences calculated on age range 18-25) (see Figure 20C) and an effect of education where participants with higher educational attainment show superior performance (see Figure 20D).

Practice effects on this test would be considerable without the establishment of alternate forms, as participants can remember previous choices. Alternate forms would protect against such effects.

### Validation

Vocabulary tests are widely used as measures of general cognitive ability and as a “hold” test or test of “premorbid iq”, since performance is relatively insensitive to variations in health in the short-term, psychological state, or many forms of brain damage (Lezak et al., 2012). Vocabulary tests provide a good control or baseline measure in populations that are reasonably well-educated and where individuals would be expected to have native (English language) fluency.

The (hard) TMB Vocabulary test correlates robustly with SAT verbal (rho(1356)=0.51, 95% CIs [0.47, 0.55]); this correlation is comparable to prior reports of correlations between well-validated vocabulary tests and SAT verbal (Mayer & Massa, 2003). As expected (Mayer & Massa, 2003; Rohde & Thompson, 2007), Vocabulary correlates to a lesser degree, but still robustly, with SAT math (rho=0.27, n=1345, 95% CIs [0.22, 0.32]) and with Matrices (rho=0.31, n=10,000, 95% CIs [0.29, 0.33]).

Controlling for participant age, TMB Vocabulary test performance (easy version; 30 item) correlates modestly with performance on the TMB Matrix Reasoning test (rho = 0.29, n = 1686, 95% CIs [0.25, 0.33]), TMB Forward Digit Span (rho = 0.25, N = 9785, 95% CIs [0.23, 0.27]), TMB Multiracial Emotion Identification Test (rho = 0.23, N = 1141, 95% CIs [0.18, 0.28]), TMB Verbal Paired Associates Test (rho = 0.33, N = 9046, 95% CIs [0.31, 0.35]) and with relatively low correlations with performance on the TMB Multiple Object Tracking test (rho = 0.11, N = 2007, 95% CIs [0.07,0.15] and TMB Simple Reaction Time test (rho = 0.15, N = 13298, 95% CIs [0.13-0.17]).

### Appropriateness for Field Test Use

Overall, the TMB Vocabulary test is brief and relatively engaging test for participants with minimal technical barriers. Practice ensures that participants know what they are supposed to do and there are minimal barriers to completion.

*Device Effects.* The TMB Vocabulary test is easy to administer across a range of device and there is very little reason to believe device characteristics would influence performance. The data are consistent with this (e.g. iPad mean = 24, SD = 5.2, N = 2727; iPhone mean = 23, SD = 5.2, N = 6107; Macintosh desktop / laptop mean = 24, SD = 4.9, N = 6937).

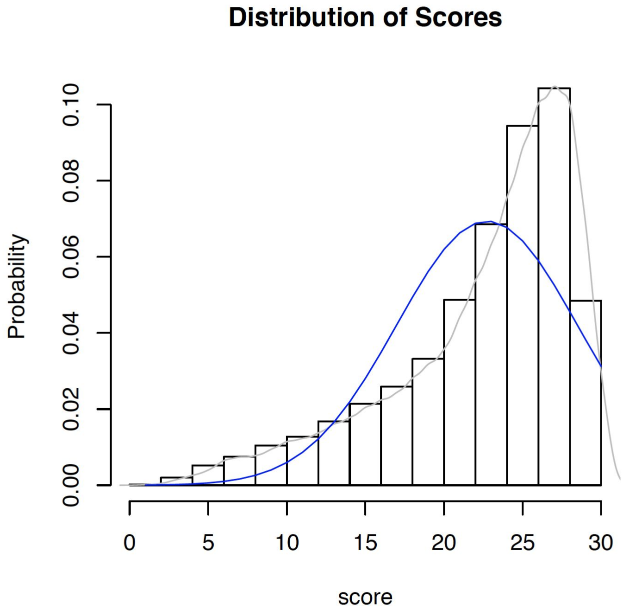
*Participant Burden.* The TMB Vocabulary test is considered enjoyable by participants and is relatively brief (4 minutes for 30 item version). Ratings on this test (3.85 / 5 stars) compare favorably with average ratings on TestMyBrain.org (3.67 / 5), with excellent completion rates compared with the rest of site (83% TMB Vocabulary vs 75% site-wide completion among consented participants).

**Further Development**

The most obvious next step for development of the TMB Vocabulary test for field test use is to create an Item Response Theory (IRT) adaptive version of the test. The independence of individual item performance combined with varying levels of difficulty mean that IRT is both appropriate for this test, and can further reduce the length of the test and eliminate ceiling effects.

Figure 20A. Distribution of Scores

EASY 30-item



HARD 20-item

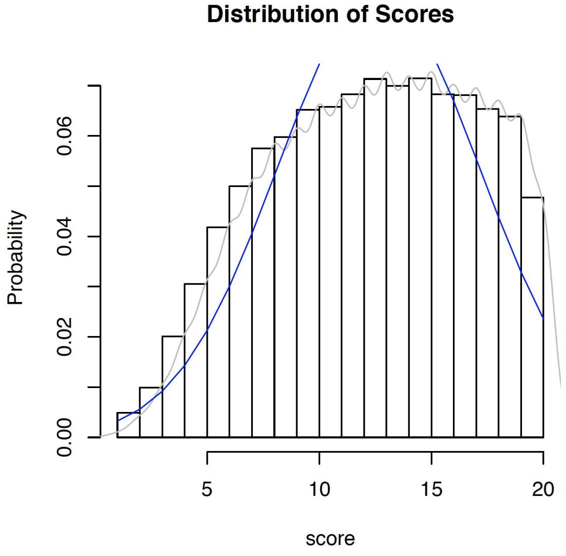
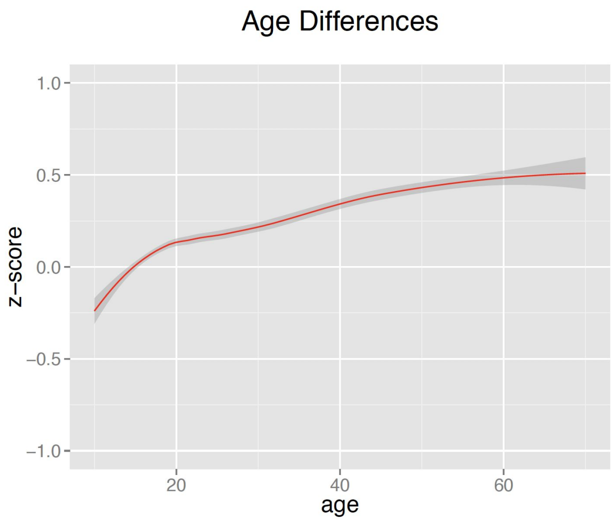


Figure 20B. Age-related Differences in Performance

EASY 30-item



HARD 20-item

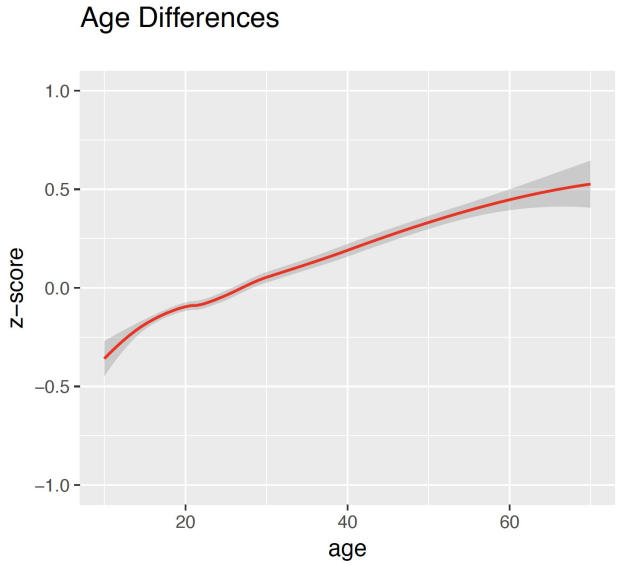
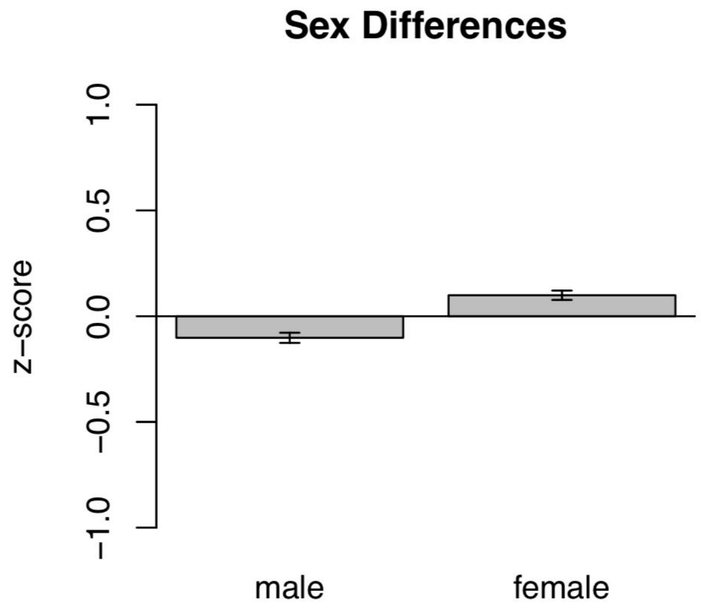


Figure 20C.  Sex Differences in Performance

EASY 30-item



HARD 20-item

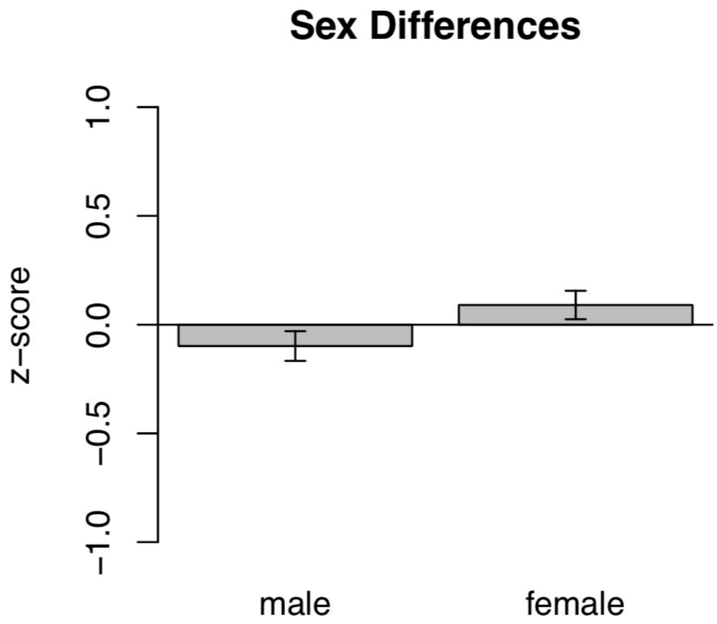
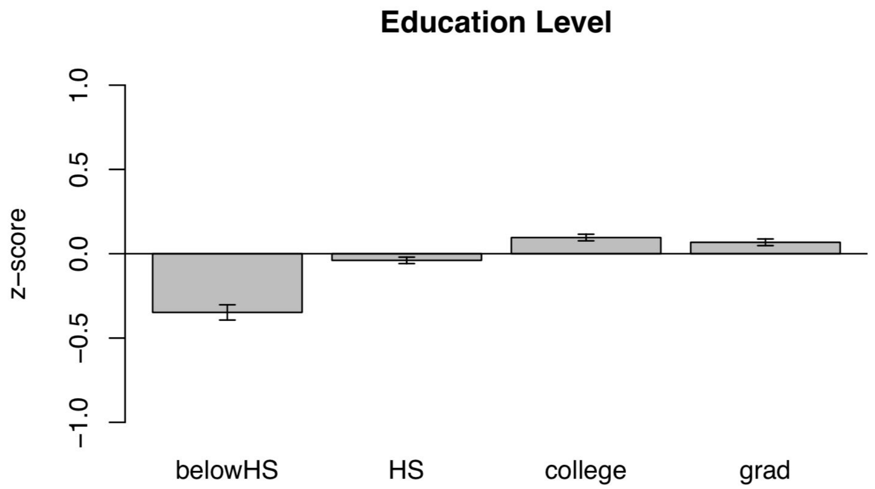
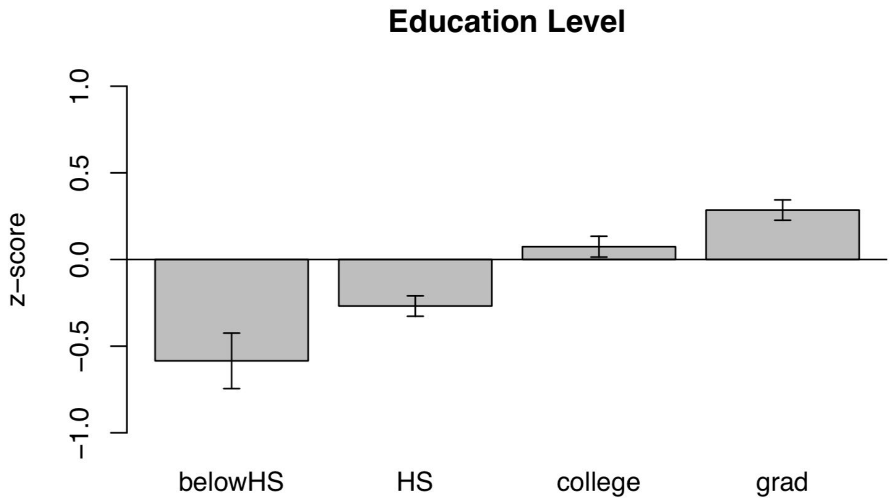


Figure 20D. Education-related Differences in Performance

EASY 30-item



HARD 20-item



## References

1. Bar-Haim, Y., Holoshitz, Y., Eldar, S., Frenkel, T.I., Muller, D., Charney, D.S., Pine, D.S., Fox, N.A., Wald, I (2010). Life-threatening danger and suppression of attention bias to threat. *Am J Psychiatry*. 167(6), 694-698.

2. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. and Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*. 42, 241-251.

3. Burt, D. B., Zembar, M. J., & Niederehe, G. (1995). Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychological Bulletin*, *117*, 285-305.

4. Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence*, 24(1), 25-52.

5. Cor, M. K., Haertel, E., Krosnick, J. A., & Malhotra, N. (2012). Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey. Social science research, 41(5), 1003-1016.

6. Deary, I., Der, G. and Ford, G. (2001). Reaction time and intelligence differences: a population based cohort study. *Intelligence*, 29(5), pp. 389-399.

7. DeGutis, J., Esterman, M., McCulloch, B., Rosenblatt, A., Milberg, W., & McGlinchey, R. (2015). Posttraumatic psychological symptoms are associated with reduced inhibitory control, not general executive dysfunction. Journal of the International Neuropsychological Society, 21(5), 342-352.

8. Dillon, D., Wiecki, T., Pechtel, P., Webb, C., Goer, F., Murray, L., . . . Pizzagalli, D. (2015). A computational analysis of flanker interference in depression. *Psychological Medicine, 45*(11), 2333-2344.

9. Dodell-Feder, D., Ressler, K. J., & Germine, L. (in press). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine.*

10. Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2012). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cerebral Cortex*, 23(11), 2712-2723.

11. Fortenbaugh, F. C., DeGutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., & Esterman, M. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological Science*, *26*(9), 1497–1510.

12. Germine, L. T., Garrido, L., Bruce, L., & Hooker, C. (2011). Social anhedonia is associated with neural abnormalities during face emotion processing. Neuroimage, 58(3), 935-945.

13. Germine, L. T., & Hooker, C. I. (2011). Face emotion recognition is related to individual differences in psychosis-proneness. Psychological Medicine, 41(5), 937-947.

14. Germine, L., Reinecke, K., & Chaytor, N.S. (2019). Digital neuropsychology: challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*.

15. Joy, S., Kaplan, E., & Fein, D. (2004). Speed and memory in the WAIS-III Digit Symbol Coding subtest across the adult lifespan. *Archives of Clinical Neuropsychology*, 19(6), 759-767.

16. Kirby, K. N., & Marakovic, N. N. (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes*, *64*(1), 22-30.

17. Lee, J. J., & Chabris, C. F. (2013). General cognitive ability and the psychological refractory period: Individual differences in the mind's bottleneck. *Psychological Science*, *24*(7), 1226–1233.

18. Lezak, M. D., Howieson, D. B., & Bigler, E. D. Tranel. D.(2012). Neuropsychological assessment, 5.

19. Maljkovic, V. & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, 22(6), 657-672.

20. Matt, G. E., Vázquez, C., & Campbell, W. K. (1992). Mood-congruent recall of affectively toned stimuli: A meta-analytic review. *Clinical Psychology Review*, *12*, 227-255.

21. Mayer, R. E., & Massa, L. J. (2003). Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. Journal of educational psychology, 95(4), 833.

22. Pizzagalli, D.A., Jahn, A.L., & O'Shea, J.P. (2005). Toward an objective characterization of an anhedonic phenotype: a signal detection approach. *Biological Psychiatry,* 57(4), 319-327.

23. Riley E., Esterman, M., Fortenbaugh, F.C., & DeGutis, J. (2017). Time-of-day variation in sustained attentional control. *Chronobiology International, 34*(7), 993-1001.

24. Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. Intelligence, 35(1), 83-92.

25. Rosenberg, M., Noonan, S., DeGutis, J. et al (2013). Sustaining visual attention in the face of distraction: A novel gradual-onset continuous performance task. *Attention, Perception, & Psychophysics*, 75(3), 426-439.

26. Rutter, L. A., Dodell-Feder, D., Vahia, I. V., Forester, B. P., Ressler, K. J., Wilmer, J. B., & Germine, L. (in press). Emotion sensitivity across the lifespan: Mapping clinical risk periods to sensitivity to facial emotion intensity. *Journal of Experimental Psychology: General*.

27. Smith, T. W., Marsden, P. V., Hout, M., & Kim, J. (2013). General Social Survey Cumulative Codebook, 1972–2012. Chicago: National Opinion Research Center.

28. Sternberg, S. (1975). Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology*, *27*(1), 1–32.

29. Vellutino F.R., Steger J.A., Harding C.J., & Phillips F. (1975). Verbal vs non-verbal paired-associates learning in poor and normal readers. *Neuropsychologia*, 13(1), 75-82.

30. Wechsler, D., & Hsiao-pin, C. (2011). WASI II: Wechsler Abbreviated Scale of Intelligence. 2nd. Psychological Corporation.

31. Wilmer, J.B., Germine, L., Ly, R., Hartshorne, J.K., Kwok, H., Pailian, H., Williams, M.A., & Halberda, J. (2012). The heritability and specificity of change detection ability. *Journal of Vision,*12(9), 1275.

32. Wilmer, J., Martini, P., Germine, L., & Nakayama, K. (2016). Multiple object tracking predicts math potential. Journal of Vision, 16(12), 421-421.